frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 29 August 2022
DOI 10.3389/fgene.2022.987238

OPEN ACCESS

EDITED BY
Ka-Chun Wong,
City University of Hong Kong, Hong Kong SAR, China

REVIEWED BY
Yaoqiang Du,
Zhejiang Provincial People's Hospital, China
Xuanbin Wang,
Hubei University of Medicine, China

*CORRESPONDENCE
Steven J. M. Jones,
sjones@bcgsc.ca

SPECIALTY SECTION
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

RECEIVED 05 July 2022
ACCEPTED 01 August 2022
PUBLISHED 29 August 2022

CITATION
Keshavarz-Rahaghi F, Pleasance E, Kolisnik T and Jones SJM (2022), A p53 transcriptional signature in primary and metastatic cancers derived using machine learning.
Front. Genet. 13:987238.
doi: 10.3389/fgene.2022.987238

A p53 transcriptional signature in primary and metastatic cancers derived using machine learning

Faeze Keshavarz-Rahaghi[1,2], Erin Pleasance[1], Tyler Kolisnik[1,3] and Steven J. M. Jones[1,4,5]*

[1]Canada's Michael Smith Genome Sciences Centre, BC Cancer, Vancouver, BC, Canada, [2]Department of Bioinformatics, University of British Columbia, Vancouver, BC, Canada, [3]School of Natural and Computational Sciences, Massey University, Auckland, New Zealand, [4]Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada, [5]Department of Molecular Biology and Biochemistry, Simon Fraser University, Vancouver, BC, Canada

The tumor suppressor gene, *TP53*, has the highest rate of mutation among all genes in human cancer. This transcription factor plays an essential role in the regulation of many cellular processes. Mutations in *TP53* result in loss of wild-type p53 function in a dominant negative manner. Although *TP53* is a well-studied gene, the transcriptome modifications caused by the mutations in this gene have not yet been explored in a pan-cancer study using both primary and metastatic samples. In this work, we used a random forest model to stratify tumor samples based on *TP53* mutational status and detected a p53 transcriptional signature. We hypothesize that the existence of this transcriptional signature is due to the loss of wild-type p53 function and is universal across primary and metastatic tumors as well as different tumor types. Additionally, we showed that the algorithm successfully detected this signature in samples with apparent silent mutations that affect correct mRNA splicing. Furthermore, we observed that most of the highly ranked genes contributing to the classification extracted from the random forest have known associations with p53 within the literature. We suggest that other genes found in this list including GPSM2, OR4N2, CTSL2, SPERT, and RPE65 protein coding genes have yet undiscovered linkages to p53 function. Our analysis of time on different therapies also revealed that this signature is more effective than the recorded *TP53* status in detecting patients who can benefit from platinum therapies and taxanes. Our findings delineate a p53 transcriptional signature, expand the knowledge of p53 biology and further identify genes important in p53 related pathways.

KEYWORDS

p53 pathway activation, transcriptome, pan-cancer, machine learining, random forest, ensemble classifier

Introduction

The most frequently somatically mutated gene in human cancer is *TP53* which encodes the p53 protein, signifying the importance of its wild-type function in tumor suppression (Levine and Oren, 2009; Kandoth et al., 2013; Duffy et al., 2014; Donehower et al., 2019; Thomas et al., 2022). Wild-type p53 functions as a transcription factor activated in response to cellular stresses (Giaccia and Kastan, 1998; Prives and Hall, 1999; Vousden and Lu, 2002; Vousden and Prives, 2009; Duffy et al., 2014; Thomas et al., 2022). The protein function can be compromised via various mechanisms (Shvarts et al., 1996; Kostic et al., 2006; Vassilev, 2007; Duffy et al., 2014), the most common being missense mutations followed by a loss of heterozygosity resulting in the complete loss of wild-type p53 function (Levine and Oren, 2009; Duffy et al., 2014; Donehower et al., 2019; Mantovani et al., 2019). Mutations in *TP53* are associated with a poor prognosis in many cancers and germline mutations in this gene cause the Li-Fraumeni syndrome which increases the susceptibility to diverse cancer types (Malkin et al., 1979; Srivastava et al., 1990; Kandoth et al., 2013; Donehower et al., 2019; Mantovani et al., 2019).

Machine learning (ML) approaches have been used to investigate large and complex data sets, including the classification of cancer types, the determination of informative features in cancer diagnosis, and the analysis of *TP53* mutations and their effects (Danziger et al., 2009; Chitrala et al., 2019; Grewal et al., 2019; Lim et al., 2019; Banerjee and Mitra, 2020; Yuan et al., 2020). Transcript expression data has been used to classify primary tumors and breast cancer subtypes based on *TP53* mutational status (Benor et al., 2020; Saghaleyni et al., 2021; Zhang et al., 2021). Subsets of The Cancer Genome Atlas (TCGA) samples have been successfully stratified based on aberrant p53 pathway activities (Zhang et al., 2021). In all of these studies, filtering and data reduction were applied to both samples and gene sets. To our knowledge, the effects of *TP53* mutations on the transcriptome have not been investigated in a pan-cancer study using both primary and metastatic samples without applying specific filters on the sample types and the genes.

In this work, we show how a transcriptional signature for loss of p53 function can be detected using ML approaches. We trained a random forest (RF) algorithm using primary tumor TCGA expression and mutation data (Kandoth et al., 2013; Weinstein et al., 2013) and metastatic tumor data from the British Columbia (BC) Cancer Agency Personalized OncoGenomics (POG) program (Pleasance et al., 2020). In this pan-cancer analysis, all coding and non-coding genes identified in both TCGA and POG datasets were included. We were able to show that our model could predict the *TP53* mutational status of tumors accurately and precisely from transcriptome expression levels alone. The list of genes contributing most to classification in the model correlated highly with those genes known to be involved in p53 function and biology. Additionally, we showed that the model could correctly categorize the samples with synonymous somatic mutations at splice sites in *TP53* as pathogenic. Our results also showed that combining all tumor types within the training set improved the overall accuracy and specificity of predictions. This indicates that a general transcriptional signature of p53 functional loss exists, is detectable and is conserved across tumor types. This signature can aid in identifying patients who can benefit from different therapies through recognition of the transcriptional patterns that are associated with p53 pathways disruptions. Due to variable response to different drug regimens, side effects, and resistance, there is a need for personalized therapies (Daly, 2010; Wilke and Dolan, 2011; Gerlinger et al., 2012; Dey et al., 2017; Hyman et al., 2017) to increase the success of treatment and improve patient outcomes especially in metastatic disease (Gerlinger et al., 2012; Dey et al., 2017).

Methods

Expression matrices and mutation data were obtained from the TCGA and POG studies (Supplementary Material). Non-primary tumor samples and the samples lacking mutation data from TCGA were excluded. All genes common to both TCGA and POG expression matrices were included. Principal Component Analysis (PCA) plots were generated (Hunter, 2007; Pedregosa et al., 2011; Waskom, 2021). Samples were divided by mutational status (mutated vs. wild-type), and further by impact of mutation (impactful vs. non-impactful). Samples with a mutation of type “silent”, “intron”, “3-prime UTR”, “5-prime UTR”, “downstream gene”, “upstream gene” or “splice region” were classified as “non-impactful”.

Random forest performance

For this analysis, only samples with “impactful” mutations or wild-type p53 copies were included to increase the likelihood of only pathogenic *TP53* driver alterations being used for training. The main hyperparameters were calibrated using 90% of samples and the obtained values were validated using the remaining 10%. The 90–10 split was performed randomly while maintaining the proportion of *TP53* mutant and wild-type samples. The RF performance was then evaluated across TCGA, POG, and merged (all TCGA and POG samples with wild-type p53 or impactful *TP53* mutations) datasets using 5-fold cross-validation analyses. Precision, recall, F1-score, area under the precision-recall curve (AUPRC), and area under the receiver operating characteristic curve (AUROC) values were found by applying the scikit-learn library functions (Pedregosa et al., 2011). To compare the performance on each cancer cohort individually versus the

pan-cancer set, the accuracy scores were compared to the Out of Bag scores which were found by training the model using only the samples in each cancer type. Additionally, performance metrics were calculated across cancer types by clustering the samples in each cohort and their predictions from the previous step and computing the values using the scikit-learn evaluation functions.

Significant genes in classification

The genes that played a more important role in classification were extracted based on the Gini importance scores of the RF model. The threshold for the number of important genes was found using a permutation-based method (Supplementary Material). The 67 genes meeting this threshold were extracted and used to perform a Gene Set Enrichment Analysis (GSEA) using the Database for Annotation, Visualization, and Integration Discovery (DAVID) (Huang et al., 2009a; Huang et al., 2009b). Cellular pathways correlated with these 67 genes were obtained with a threshold of 0.05 for p-values adjusted using the Benjamini–Hochberg procedure (Huang et al., 2009b).

Prediction probabilities and outliers

The probabilities associated with the RF predictions were extracted from the model and were grouped by prediction correctness, sample source (TCGA and POG), true *TP53* status (label), and the *TP53* status predicted by the RF. Using these likelihoods, mispredicted samples with high prediction probabilities (>0.95) were identified. Two samples (TCGA-AR-A24T-01 and TCGA-VM-A8CH) that belonged to relatively balanced cancer cohorts were investigated. Whole exome sequencing and RNA-seq files of these samples and RNA-seq files of four other comparator brain and breast cancer samples (two with wild-type and two with mutated *TP53* copies) were visualized using the Broad Institute's Integrative Genomics Viewer (IGV) (Robinson et al., 2011).

Samples with non-impactful *TP53* mutations

To determine the status of the samples with non-impactful mutations, the merged set was used to train the algorithm, and the status of the non-impactfully mutated samples were predicted. The samples with silent mutations assigned to the p53 mutant category were further investigated, as the expectation would be to see no change in the p53 protein and therefore similar behavior to the p53 wild-type category. The RNA-seq data of these samples was visualized in IGV (Robinson et al., 2011).

Treatment efficacy in patients with mutated and wild-type p53

Treatment data from the POG cohort was obtained (Pleasance et al., 2020), and drugs were grouped by their mechanism of action and/or target genes or proteins (Antineoplastic Agents, 2012; Vardanyan et al., 2016; Wishart et al., 2018; NCI, 2022). Combination therapies were separated into individual drugs, and data for patients on a double-blind trial where the received treatment was unknown were filtered out. Total days on therapy was used as a proxy for treatment response as response data was not available. Drug groups with <5 patients or only p53 wild-type tumors were excluded.

Results

Primary tumors from TCGA with available mutation data and metastatic tumors from POG were collated, and 48,784 overlapping coding and non-coding genes were identified. The PCA indicated that samples clustered by cancer type (Supplementary Figures S1A,B) and primary and metastatic samples also clustered together (Supplementary Figure S1C). Since there was no distinctive boundary between TCGA and POG data sets, the samples were merged for classification.

Out of the 8755 TCGA samples, 3,373 (39%) had a mutation in *TP53* and 5,382 (61%) had only wild-type copies of this gene. 47 (1%) of the samples with a mutated *TP53* were classified as "non-impactful" while the other 3,326 (99%) were placed into the "impactful" category. Out of 570 POG samples, 229 (40%) had a mutation in *TP53* and 341 (60%) contained wild-type copies. Among 229 mutated samples, 23 (10%) were categorized as "non-impactful", and 206 (90%) as "impactful".

Random forest performance

The performance of the RF was first optimized by tuning the hyperparameters for the TCGA, POG, and merged data sets and then evaluated using 5-fold cross-validation analyses (Supplementary Table S1 and Supplementary Figures S2–S5). Over 10 independent tests, merging TCGA and POG resulted in a mean of 35 more samples (4 TCGA and 31 POG) successfully classified by the RF, demonstrating benefit in combining the primary and metastatic samples to train the algorithm and detect the transcriptional patterns. Overall, the RF performance was successful with AUROC 0.94 and AUPRC 0.96 in the merged dataset (Supplementary Table S2 and Figure 1). Mean accuracy was 0.88, with precision 0.88, recall 0.87, and f1-score 0.87.

Across the TCGA cancer types, the algorithm classified samples with >0.75 f1-score in most cancer types where the

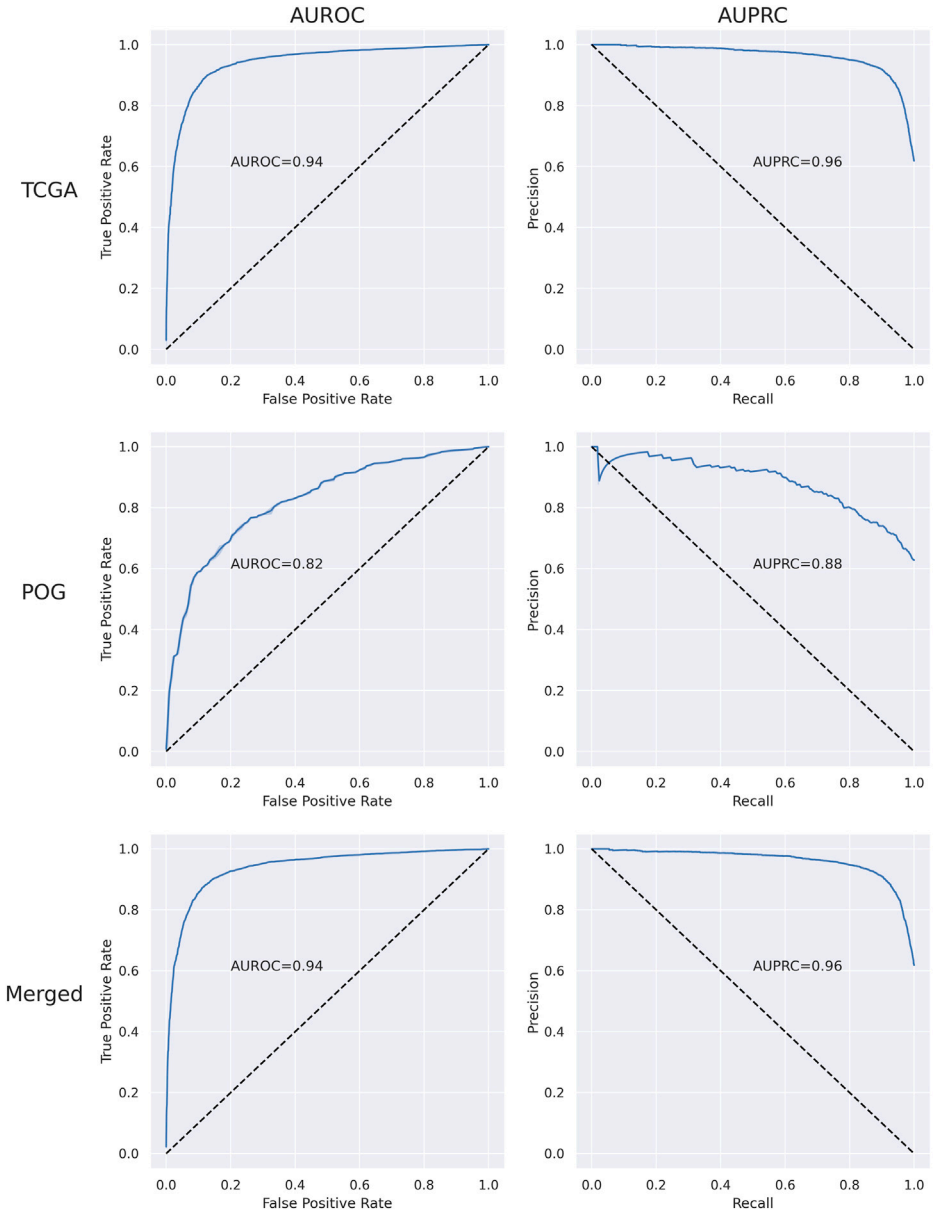

FIGURE 1
Plots of area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC) for TCGA, POG, and merged (set of all impactful and wild-type samples) data sets.

minority class to majority class sample number ratio was greater than 10% (Supplementary Table S3). In highly imbalanced cancer types, sample classification was less successful. The prediction accuracy when all the samples were used for training was better than or very similar to when training was done on individual cancer types for 30 out of 33 tumor categories (Supplementary Figure S6), so it can be concluded that there is a benefit in combining all cancer types to train the RF model.

Significant genes in classification

To understand the underlying decision-making process of the RF algorithm, the genes that played a more important role in classification were extracted using the built-in feature importance scores (Gini importance scores). A threshold of 67 genes was attained using a permutation method (Supplementary Material) as the cut-off for the list of genes contributing to the classification of samples based on *TP53* mutation status (Supplementary Figure S7). The importance scores of these genes as well as

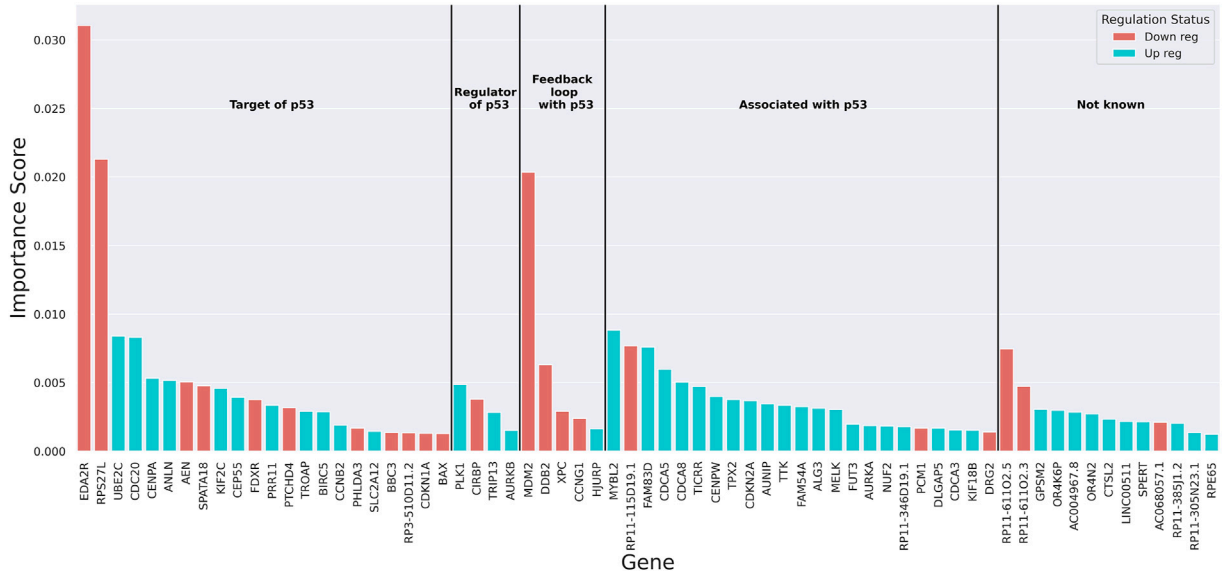

FIGURE 2
The bar plot of importance (Gini) scores of top 67 genes extracted from the random forest. Red colour indicates downregulation and blue colour shows upregulation of the genes with the loss of wild-type p53 function. Genes are grouped based on their known link to p53.

the change in their expression with the loss of wild-type p53 function and their known link to p53 based on retrieved literature were obtained (Figure 2 and Supplementary Figure S4). Over- and under-expression of these genes in the absence of wild-type p53 function was observed to be entirely aligned with what is known about the regulation of these genes (Supplementary Figure S4). For example, the three genes with the highest importance scores in the classification have all been shown to be upregulated in the presence of wild-type p53 activity (Piette et al., 1997; Tanikawa et al., 2010; Xiong et al., 2011; Donehower et al., 2019; Moyer et al., 2020). The findings from other studies on p53 targets using experimental approaches further confirmed the relevance of our gene list. Out of the 27 genes known to be regulated by p53 in our list (targets of p53 and genes that are in a feedback loop with p53), 10 were found in the list of 122 p53-regulated genes by Riley et al. (p-value = 8.9×10^{-16} from a hypergeometric test with N = 18,337 (protein-coding genes in our data), k = 122, n = 27, and x = 10) (Riley et al., 2008). Additionally, 9 out of the 27 genes regulated by p53 were found in the list of 46 genes bound by p53 identified by Nikulenkov et al. (p-value = 7.7×10^{-18} from a hypergeometric test with N = 18,337, k = 46, n = 27, and x = 9) (Nikulenkov et al., 2012). Expression of our model's top 10 genes (Supplementary Figure S8) confirmed an overall difference in the transcriptional behavior of these genes between the mutant and wild-type classes. GSEA revealed these 67 genes were most enriched in cell cycle pathways ($P_{adj} < 8.9 \times 10^{-22}$), which are expected to be affected by *TP53* mutations (Huang et al., 2009a; Huang et al., 2009b) (Supplementary Table S5). Lastly, to recognize the potential role of the *TP53* transcript itself in the classification, the rank of *TP53* in the model's important genes list was obtained in 100 independent training iterations. This rank on average was 109 (±6 standard deviation) with a median of 109 (IQR: 105–114), indicating that the transcriptional level of the *TP53* gene itself did not significantly contribute to the classification.

Prediction probabilities and outliers

Correctly predicted samples had a higher median prediction probability, indicating higher confidence compared to mis-predicted samples ($P_{adj} < 1 \times 10^{-4}$) (Supplementary Figure S9). Prediction probabilities for TCGA were higher than for POG, likely due in part to the much larger number of samples included in training the algorithm. Furthermore, it was noted that the wild-type samples for p53 had higher probabilities compared to samples with mutated *TP53* genes. This could be due to the larger number of wild-type p53 samples in the training set or could indicate that there is a more dominant signature of wild-type p53 function across different cancer types.

Out of 1,402 samples with a prediction probability of more than 0.95, 14 (1%) were found to be incorrectly classified (Supplementary Table S6). The majority of these samples belonged to highly imbalanced cancer cohorts. However, two samples belonged to relatively balanced cohorts (TCGA-VM-A8CH-01, brain lower grade glioma, and TCGA-AR-A24T-01, breast invasive carcinoma). Inspection of the RNA-seq

alignments revealed that the TCGA-VM-A8CH-01 contained a missense mutation at the 277th amino acid of p53 that changes cysteine to phenylalanine (p.C277F) (Supplementary Figure S10A). Whole exome sequencing data showed that both tumor and normal samples of this patient contained this single nucleotide variation and thus the mutation was germline (the sample was initially mislabeled as wild-type due to lack of a somatic mutation). Based on the *TP53* Database (R20 July 2019: https://tp53.isb-cgc.org) (Bouaoun et al., 2016), it is believed that p. C277F is a pathogenic, non-functional mutation (i.e., a loss-of-function mutation). Our findings confirm the loss of p53 activity and suggest that this mutation might play a role in cancer predisposition. The label of this lower grade glioma sample was then changed to p53 mutant, and the algorithm performance was inspected again. No change was observed in the performance metrics which shows that the RF model is already robust to noise.

The whole exome sequencing and RNA-seq data for TCGA-AR-A24T-01 breast invasive carcinoma confirmed the existence of a p. R273H mutation in *TP53* (Supplementary Figure S10B) even though it was classified as wild-type tumor by the RF. The misprediction does not seem to be related to the specific mutation because out of the 101 TCGA and POG samples with the p. R273H mutation in our data, 89 (88%) were correctly assigned to p53 mutant category. The mean of the prediction probability associated with the 11 mispredicted samples with p. R273H mutation (excluding TCGA-AR-A24T-01 sample) was 0.67 (±0.12 sd) which is considerably lower than the prediction probability of TCGA-AR-A24T (0.96). To determine if this is related to clonality or tumor content, we looked at the variant allele frequency (VAF) of all the samples containing a p. R273H mutation. The average VAF of the 89 correctly classified samples was 0.56 (±0.20 sd) with a median of 0.55 (IQR: 0.39–0.71) while the average VAF of the 11 misclassified samples was 0.31 (±0.16 sd) with a median of 0.26 (IQR: 0.19–0.44). The VAF of TCGA-AR-A24T-01 was approximately 0.30 which is closer to the mean and median of the mispredicted samples. Considering the low VAF, it is possible that low tumor content in this sample might account for the incorrect prediction.

Samples with non-impactful *TP53* mutations

The samples with non-impactful mutations were excluded from all the previous analyses due to ambiguity in their pathogenicity. To discern the effect of non-impactful mutations, the algorithm was trained on the merged set, and the mutational status of the samples with non-impactful *TP53* mutations was determined by the fully trained RF (Supplementary Table S7). In most mutation groups, many of the samples were assigned to the wild-type category as expected, whereas 30 out of 38 (80%) of the silent mutations were categorized as p53 mutant. All these silent mutations have been previously reported in patients with cancer, Li-Fraumeni syndrome, or other conditions related to cancer based on the NCBI ClinVar database (Table 1) (Landrum et al., 2018). The c.375G>T, c.375G>A, and c.375G>C (p.T125T) mutations occur at the last nucleotide of exon 4 and were shown to disrupt the *TP53* mRNA splicing either through molecular studies or splice site predictive tools (NM_000546, 2379; NM_000546, 1778; NM_000546a). The c.672G>A (p.E224E) mutation occurs at the last nucleotide of exon 5 and was shown to lead to aberrant mRNA splicing (NM_000546b). The c.993G>A (p.Q331Q) mutation is located at the last nucleotide of exon 8 and is predicted to affect the normal mRNA splicing (NM_000546, 2886). Supek et al. have also demonstrated that p. T125T, p. E224E, and p. Q331Q mutations are enriched in *TP53* and suggested that they have a functional role in cancer (Supek et al., 2014).

The RNA-seq data confirms that the silent mutations which occur at the end of exons (p.T125T, p. E224E, and p. Q331Q) affect the mRNA splicing in these samples (Figure 3 and Supplementary Figure S11). Introns 4, 5, and 8 were not successfully spliced out in samples bearing p. T125T, p. E224E, and p. Q331Q mutations respectively. The only exception to the association between these silent mutations and intron retention was for TCGA-CR-7401-01 where the p. E224E mutation did not appear in the RNA-seq data and splicing appeared normal (Supplementary Figure S11C). With a VAF of 0.14, it may likely be subclonal and the expression not detected. The RNA-seq data of the sample with p. A69A appeared indistinguishable from the other lung squamous cell carcinoma sample with wild-type p53 (Supplementary Figure S11E). This is consistent with its classification in ClinVar as likely benign (NM_000546, 2198). Further investigation is needed to understand why this sample was assigned to the p53 mutant category by the algorithm.

Treatment efficacy in patients with mutated and wild-type p53

We sought to explore whether *TP53* mutation status was predictive of therapy response for patients within the POG cohort for 29 drug groups (Supplementary Table S8). A longer time on therapy was interpreted to be indicative of ongoing clinical benefit for the patients (Pleasance et al., 2020). Time on therapy was more strongly associated with *TP53* status predicted by the RF than with recorded *TP53* mutation status for platinum therapies (Bonferroni-adjusted pvalue 0.001 vs. 0.027) and taxanes (0.041 vs. 0.288), with longer duration in predicted *TP53*-mutant cases (Figure 4). The majority of these therapies (94% of platinums and 77% of taxanes) were received in combination with other drugs. The reverse association is true for the drug group Epothilones (represented only by the drug eribulin), where the recorded *TP53* mutation status was more strongly associated with time on therapy (Bonferroni-adjusted *p*-value 0.138 vs. 0.014), with longer

TABLE 1 Silent mutations classified as p53 mutant, the number of samples containing these mutations, and the reported consequences and interpretation of them based on ClinVar database (nucleotide variations with * are not present in general population based on ClinVar evidence).

Nucleotide variation	Amino acid variation	Number of samples	Exon location	Disease	ClinVar pathogenicity	ClinVar record
c.375G>T*	p.T125T	20	Last nucleotide of exon 4	Li-Fraumeni syndrome	Likely pathogenic	VCV000237948.3 (NM_000546, 2379)
c.375G>A*	p.T125T	3	Last nucleotide of exon 4	Li-Fraumeni syndrome Li-Fraumeni-like/Chompret criteria Rhabdomyosarcoma Breast and/or ovarian cancer Malignant tumour of prostate	Pathogenic	VCV000177825.18 (NM_000546, 1778)
c.375G>C	p.T125T	1	Last nucleotide of exon 4	Ependymoma Early-onset breast cancer	Likely pathogenic	VCV000480746.3 (NM_000546a)
c.672G>A*	p.E224E	3	Last nucleotide of exon 5	Li-Fraumeni syndrome Chompret criteria	Pathogenic/Likely pathogenic	VCV000080709.6 (NM_000546b)
c.993G>A*	p.Q331Q	2	Last nucleotide of exon 8	Adrenocortical carcinoma Suspected Li-Fraumeni syndrome	Likely pathogenic	VCV000428868.7 (NM_000546, 2886)
c.207T>C	p.A69A	1	Exon 4	Li-Fraumeni syndrome Hereditary cancer-predisposing syndromes	Likely benign	VCV000219841.7 (NM_000546, 2198)

treatment duration in predicted *TP53*-wild-type cases (Figure 4). For the remaining drug categories, the classification of data points by the recorded *TP53* status and the RF predictions were statistically similar (Supplementary Figure S12).

Discussion

We developed a RF model derived from tumor transcriptomes which can successfully classify primary and metastatic tumor samples based on *TP53* aberrations. We hypothesize that a specific transcriptional signature is associated with the loss of functional p53 within mutated tumors regardless of their type or tissue of origin since the algorithm stratified 88% of the merged set samples correctly, and the combination of different cancer types as well as the inclusion of both primary and metastatic sets during training enhanced the performance of the model. The TCGA primary and POG metastatic sets were combined for training to discover potential differences in p53 biology and the cellular pathways associated with this transcription factor between the two sets, however, the results from the merged set showed that the p53 inactivation signature is universal across all tumors. We used this p53 transcriptional signature to identify the genes with core roles in p53 processes, determine the functional relevance of silent mutations, and better predict response to treatment.

The choice of the RF algorithm for this work is due to its ability to find complicated patterns in data and improve classification with less overfitting compared to other models (Breiman, 2001; Khoshgoftaar et al., 2013; Maurya et al., 2021). RFs have also been shown to be robust to noise and perform better on imbalanced data sets (Breiman, 2001; Khoshgoftaar et al., 2007; Khoshgoftaar et al., 2013). Even when including highly imbalanced cohorts, the RF model had better performance metrics than the published XGBoost model used to classify p53 pathway activity (Zhang et al., 2021). The RF could additionally identify samples with germline *TP53* mutations and samples with silent mutations that affect the mRNA splicing. These findings confirm the existence of a p53 transcriptional signature and the power of the RF algorithm to detect this signature. The RF, combined with drug treatment data, revealed that the presence of the mutant *TP53* signature was associated with a longer time on therapy for platinum and taxane therapies. It is important to highlight that most patients in the POG cohort received combination therapies, and treatment effectiveness has been shown to be affected by the mode of therapy (Murray and Mirzayans, 2020; Shu et al., 2022). Moreover, our treatment data set was relatively small, so further work will be needed to confirm the treatment efficacy results presented here.

The majority of significant genes in the classification found in this work have known links with p53 function. Based on this strong association, we speculate that many of the other significant transcripts possess an unappreciated role in p53 biology; these include protein coding genes GPSM2, OR4N2, CTSL2, SPERT, and RPE65; pseudogenes RP11-611O2.3, OR4K6P and AC004967.8; long non-coding RNAs (lncRNAs) RP11-611O2.5, LINC00511, AC068057.1, RP11-385J1.2, and RP11-305N23.1. There is existing evidence for roles in tumor biology for many of these genes. GPSM2 has a role in

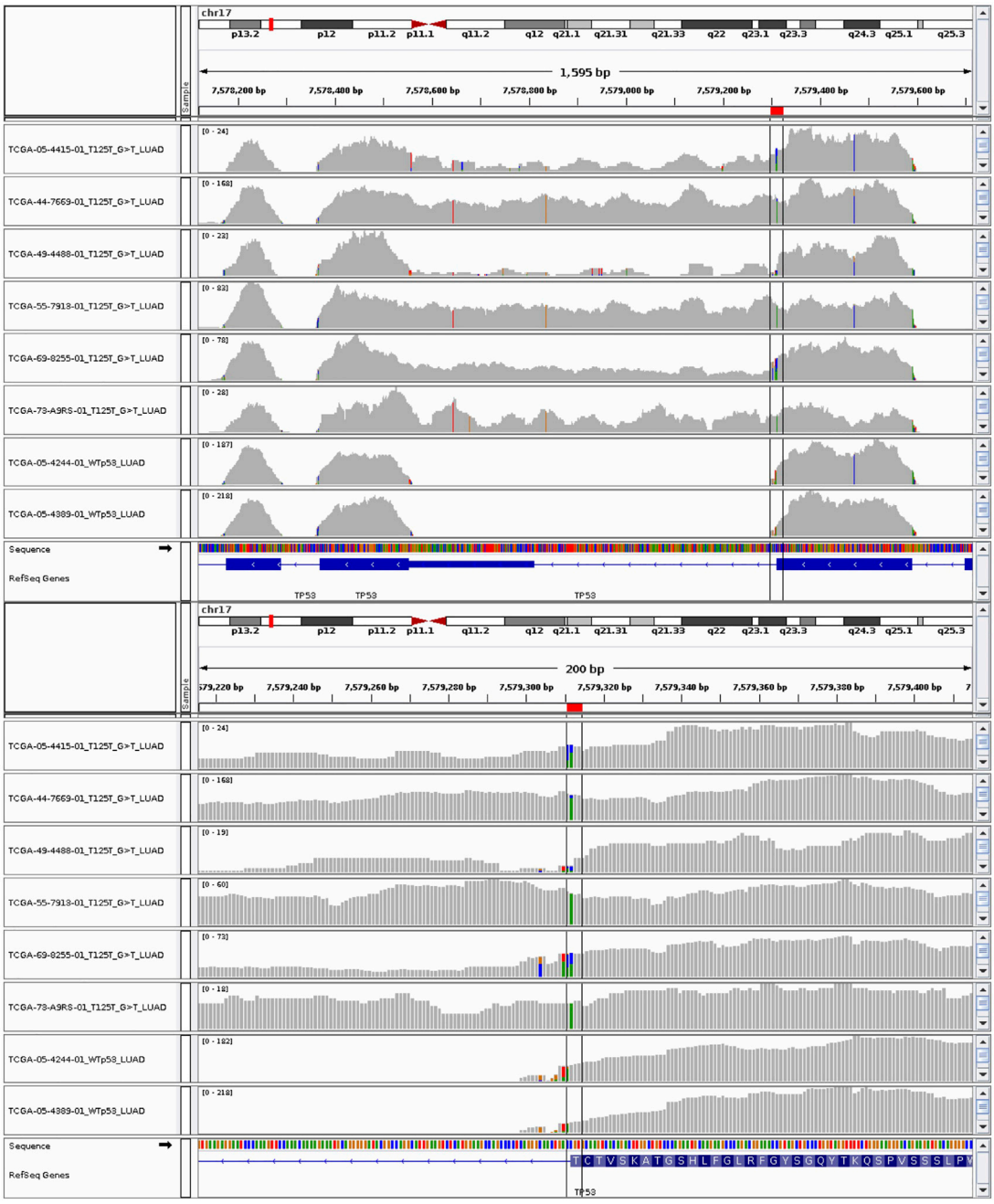

FIGURE 3
RNA-seq data of lung adenocarcinoma (LUAD) samples with p53 silent mutations at threonine 125 with specific nucleotide modification of G>T. The last two tracks are from LUAD samples with wild-type p53 copies.

breast cancer cell division and promotes tumor proliferation and metastasis in hepatic cellular cancer (Blumer et al., 2006; Fukukawa et al., 2010; He et al., 2017). Additionally, it has been shown that GPSM2 plays an important role in mitosis (Fukukawa et al., 2010) and our GSEA showed that mitosis and cell division pathways are among the top pathways found to be significant in classification. OR4N2 was shown to be mutated on at least two sites in epithelial ovarian cancers (Zhang et al., 2019). OR4N2 encodes a G protein-coupled receptor and GPSM2 participates in activation of G proteins (Blumer et al., 2006; Maßberg and Hatt, 2018). The p53 signaling pathway, which is the eighth important pathway in our classification, contains several G protein-coupled interactions which highlights the importance of these

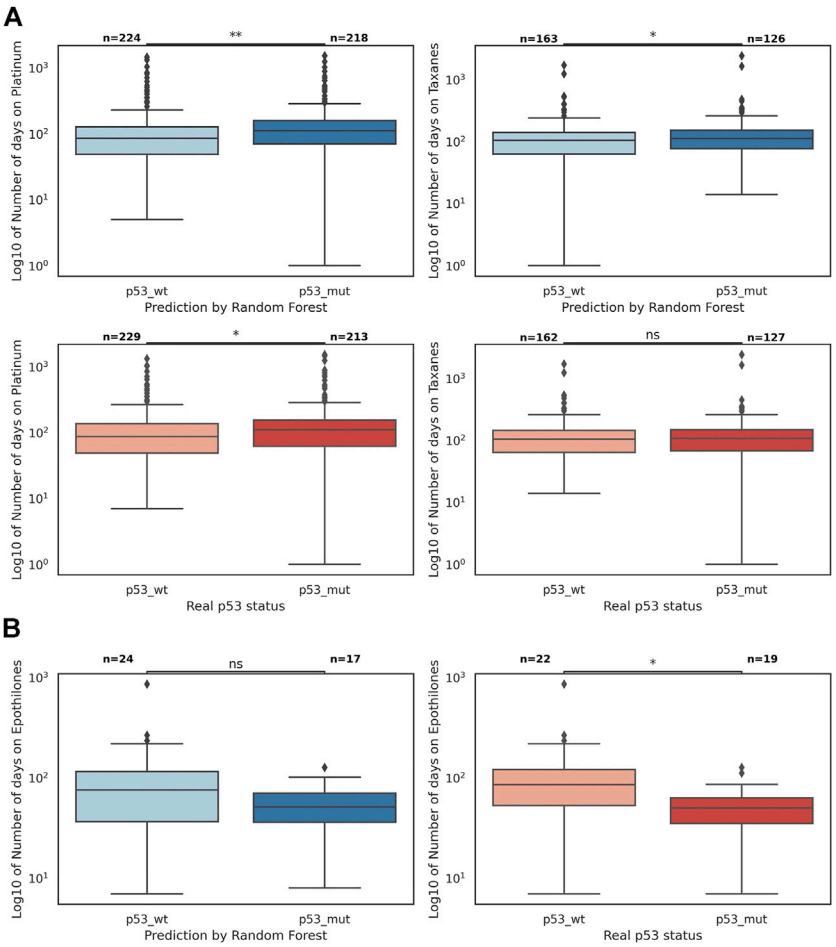

FIGURE 4
The number of days on platinum therapies, taxanes, and epothilones divided by *TP53* mutation status and the predicted status by the random forest (the *p*-values are found in a Mann-Whitney-Wilcoxon two-sided test with Bonferroni correction; *p*-value annotation guide: ns: 5.00e-02 < p ≤ 1.00, *: 1.00e-02 < p ≤ 5.00e-2, **: 1.00e-03 < p ≤ 1.00e-02). **(A)** The boxplots of log10 of the number of days on platinum and taxanes; the difference between p53 wild-type and mutant sets is statistically more significant when data points are divided by the random forest predictions (blue) than when they are divided by the true p53 status (red). **(B)** The boxplots of log10 of the number of days on epothilones (represented only by the drug eribulin); the difference between p53 wild-type and mutant sets is statistically more significant when data points are divided by the true p53 status (red) than when they are divided by the random forest predictions (blue).

genes in p53 function (Solyakov et al., 2009). CTSL2 was demonstrated to be highly expressed in various human cancers and was speculated to be associated with metastasis (Santamaría et al., 1998; Liu et al., 2004; Sun et al., 2016). Knockdown of SPERT was shown to lead to tumor growth suppression and apoptosis (Zheng and Chen, 2018). RPE65 was demonstrated to be highly downregulated in melanoma and squamous cell carcinoma of skin (Hinterhuber et al., 2005; Hassel et al., 2013). We also observed that RP11-611O2.3 and RP11-611O2.5 are located at the 3' end of the MDM2 gene and their low expression in the p53 mutant tumors is consistent with the observed low MDM2 expression in such tumors.

TP53 functions as a homo-tetramer and the inclusion of mutant protein products provides the mechanism by which p53 mutants can function in a dominant negative manner (Ko and Prives, 1996; de Vries et al., 2002; Donehower et al., 2019; Thomas et al., 2022). For tumor types where the class size was sufficient to allow robust training, over 90% of tumors exhibiting a strong p53 transcriptional signature with a likelihood of >0.75 were found to have a corresponding *TP53* mutation (95, 96, 97, and 91% of tumors respectively for breast invasive carcinoma, colon adenocarcinoma, brain lower grade glioma and lung adenocarcinoma). This indicates that there are no other gene mutations that can generate this signature at a high frequency even though mutations in other genes within the same pathway as *TP53* might have been expected to generate the same DNA repair defect phenotype and transcriptional signature. This further confirms the unique role of p53 as a key contributor to human cancer.

The use of ML in this work led to the discovery of complicated patterns in the transcriptome that otherwise could not be possible to detect. ML approaches were shown to have the ability to detect complex relationships in a fast and accurate manner in many different areas of omics sciences (Wu and Wang, 2018; Smith et al., 2020; Reel et al., 2021). These algorithms are constantly being improved and can be easily automated. Furthermore, RF models have the capability to distinguish different roles that genes might play in different cells by utilizing them at different depths of decision trees. It has been demonstrated that some genes might play different roles based on the cell type or the biological context. For example, it has been postulated that the function of *MELK* might be context dependent and can positively or negatively regulate p53 in different cell types (Seong and Ha, 2012; Gu et al., 2013; Ganguly et al., 2015). RF models can capture such context-dependent relationships since they can use the same gene at different depths of decision trees with different thresholds to split the samples.

In conclusion, we have successfully showed that a RF model can classify tumor samples based on *TP53* status regardless of their type or tissue of origin using expression data alone. The genes contributing to the signature provide insight to p53 biology, and the use of this signature for classification has the potential to aid in treatment management and identification of the patients who can benefit from therapies related to *TP53* status.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: tcga_RSEM_gene_tpm was downloaded from https://xenabrowser.net/on 21 June 2021 and the MC3 Public MAF mutation data (mc3. v0.2.8. PUBLIC.maf. gz) was downloaded from: https://gdc.cancer.gov/about-data/publications/mc3-2017 on 30 September 2020.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

Author contributions

SJ and FK-R designed and conceptualized the project. FK-R and EP acquired the relevant data. FK-R performed all the analyses, generated all the figures, and drafted the manuscript. SJ and FK-R interpreted the results. SJ, EP, TK, and FK-R added relevant revisions to the manuscript. SJ supervised the project and obtained funding. All authors approved the final manuscript.

Funding

This research has been supported by the BC Cancer Foundation, Canada Research Chair Program funding to SJ and a University of British Columbia's 4-Year Doctoral Fellowship Award to FK.

Acknowledgments

This work would not be possible without the participation of our patients and families, the POG team, the GSC platform, the GSC Bioinformatics, Systems, and Data Management teams, and the generous support of the BC Cancer Foundation and Genome British Columbia (project B20POG). We also acknowledge contributions towards equipment and infrastructure from Genome Canada and Genome BC (projects 202SEQ, 212SEQ, 262SEQ, 12002), Canada Foundation for Innovation (projects 20070, 30981, 30198, 33408 and 35444) and the BC Knowledge Development Fund. Other contributors include Roche Pharma. The results published here are in part based upon data generated by the following projects and obtained from dbGaP (http://www.ncbi.nlm.nih.gov/gap): The Cancer Genome Atlas managed by the NCI and NHGRI (http://cancergenome.nih.gov); Genotype-Tissue Expression (GTEx) Project, supported by the Common Fund of the Office of the Director of the National Institutes of Health (https://commonfund.nih.gov/GTEx).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.987238/full#supplementary-material

References

Antineoplastic Agents [Internet]. LiverTox: Clinical and Research Information on Drug-Induced Liver Injury. Bethesda (MD): National Institute of Diabetes and Digestive and Kidney Diseases; 2012 [cited 2022 Jan 24]. Available from: https://www.ncbi.nlm.nih.gov/books/.

Banerjee, A., and Mitra, P. (2020). Estimating the effect of single-point mutations on protein thermodynamic stability and analyzing the mutation landscape of the p53 protein. *J. Chem. Inf. Model.* 60 (6), 3315–3323. doi:10.1021/acs.jcim.0c00256

Benor, G., Fuks, G., Chin, S. F., Rueda, O. M., Mukherjee, S., Arandkar, S., et al. (2020). Transcriptional profiling reveals a subset of human breast tumors that retain wt TP53 but display mutant p53-associated features. *Mol. Oncol.* 14 (8), 1640–1652. doi:10.1002/1878-0261.12736

Blumer, J. B., Kuriyama, R., Gettys, T. W., and Lanier, S. M. (2006). The G-protein regulatory (GPR) motif-containing Leu-Gly-Asn-enriched protein (LGN) and Gialpha3 influence cortical positioning of the mitotic spindle poles at metaphase in symmetrically dividing mammalian cells. *Eur. J. Cell Biol.* 85 (12), 1233–1240. doi:10.1016/j.ejcb.2006.08.002

Bouaoun, L., Sonkin, D., Ardin, M., Hollstein, M., Byrnes, G., Zavadil, J., et al. (2016). TP53 variations in human cancers: New lessons from the IARC TP53 database and genomics data. *Hum. Mutat.* 37 (9), 865–876. doi:10.1002/humu.23035

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi:10.1023/a:1010933404324

Chitrala, K. N., Nagarkatti, M., Nagarkatti, P., and Yeguvapalli, S. (2019). Analysis of the TP53 deleterious single nucleotide polymorphisms impact on estrogen receptor alpha-p53 interaction: A machine learning approach. *Int. J. Mol. Sci.* 20 (12), E2962. doi:10.3390/ijms20122962

Daly, A. K. (2010). Pharmacogenetics and human genetic polymorphisms. *Biochem. J.* 429 (3), 435–449. doi:10.1042/BJ20100522

Danziger, S. A., Baronio, R., Ho, L., Hall, L., Salmon, K., Wesley Hatfield, G., et al. (2009). Predicting positive p53 cancer rescue regions using Most Informative Positive (MIP) active learning. *PLoS Comput. Biol.* 5 (9), e1000498. doi:10.1371/journal.pcbi.1000498

de Vries, A., Flores, E. R., Miranda, B., Hsieh, H. M., van Oostrom, C. T. M., Sage, J., et al. (2002). Targeted point mutations of p53 lead to dominant-negative inhibition of wild-type p53 function. *Proc. Natl. Acad. Sci. U. S. A.* 99 (5), 2948–2953. doi:10.1073/pnas.052713099

Dey, N., Williams, C., Leyland-Jones, B., and De, P. (2017). Mutation matters in precision medicine: A future to believe in. *Cancer Treat. Rev.* 55, 136–149. doi:10.1016/j.ctrv.2017.03.002

Donehower, L. A., Soussi, T., Korkut, A., Liu, Y., Schultz, A., Cardenas, M., et al. (2019). Integrated analysis of TP53 gene and pathway alterations in the cancer genome atlas. *Cell Rep.* 28 (5), 1370–1384.e5. e5. Available from. doi:10.1016/j.celrep.2019.07.001

Duffy, M. J., Synnott, N. C., McGowan, P. M., Crown, J., O'Connor, D., and Gallagher, W. M. (2014). P53 as a target for the treatment of cancer. *Cancer Treat. Rev.* 40 (10), 1153–1160. Available from. doi:10.1016/j.ctrv.2014.10.004

Fukukawa, C., Ueda, K., Nishidate, T., Katagiri, T., and Nakamura, Y. (2010). Critical roles of LGN/GPSM2 phosphorylation by PBK/TOPK in cell division of breast cancer cells. *Genes Chromosom. Cancer* 49, 861–872. doi:10.1002/gcc.20795

Ganguly, R., Mohyeldin, A., Thiel, J., Kornblum, H. I., Beullens, M., and Nakano, I. (2015). MELK—a conserved kinase: Functions, signaling, cancer, and controversy. *Clin. Transl. Med.* 4 (1), 11. doi:10.1186/s40169-014-0045-y

Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., et al. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* 366, 883–892. doi:10.1056/NEJMoa1113205

Giaccia, A. J., and Kastan, M. B. (1998). The complexity of p53 modulation: Emerging patterns from divergent signals. *Genes Dev.* 12 (19), 2973–2983. doi:10.1101/gad.12.19.2973

Grewal, J. K., Tessier-Cloutier, B., Jones, M., Gakkhar, S., Ma, Y., Moore, R., et al. (2019). Application of a neural network whole transcriptome-based pan-cancer method for diagnosis of primary and metastatic cancers. *JAMA Netw. Open* 2 (4), e192597. doi:10.1001/jamanetworkopen.2019.2597

Gu, C., Banasavadi-Siddegowda, Y. K., Joshi, K., Nakamura, Y., Kurt, H., Gupta, S., et al. (2013). Tumor-specific activation of the C-JUN/MELK pathway regulates glioma stem cell growth in a p53-dependent manner. *Stem Cells* 31 (5), 870–881. doi:10.1002/stem.1322

Hassel, J. C., Amann, P. M., Schadendorf, D., Eichmüller, S. B., Nagler, M., and Bazhin, A. V. (2013). Lecithin retinol acyltransferase as a potential prognostic marker for malignant melanoma. *Exp. Dermatol.* 22 (11), 757–759. doi:10.1111/exd.12236

He, X. Q., Zhang, Y. F., Yu, J. J., Gan, Y. Y., Han, N. N., Zhang, M. X., et al. (2017). High expression of G-protein signaling modulator 2 in hepatocellular carcinoma facilitates tumor growth and metastasis by activating the PI3K/AKT signaling pathway. *Tumour Biol.* 39 (3), 1010428317695971. doi:10.1177/1010428317695971

Hinterhuber, G., Cauza, K., Dingelmaier-Hovorka, R., Diem, E., Horvat, R., Wolff, K., et al. (2005). Expression of RPE65, a putative receptor for plasma retinol-binding protein, in nonmelanocytic skin tumours. *Br. J. Dermatol.* 153 (4), 785–789. doi:10.1111/j.1365-2133.2005.06769.x

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 37 (1), 1–13. doi:10.1093/nar/gkn923

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4 (1), 44–57. doi:10.1038/nprot.2008.211

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* 9 (3), 90–95. doi:10.1109/mcse.2007.55

Hyman, D. M., Taylor, B. S., and Baselga, J. (2017). Implementing genome-driven oncology. *Cell* 168 (4), 584–599. doi:10.1016/j.cell.2016.12.015

Kandoth, C., McLellan, M. D., Vandin, F., Ye, K., Niu, B., Lu, C., et al. (2013). Mutational landscape and significance across 12 major cancer types. *Nature* 502 (7471), 333–339. doi:10.1038/nature12634

Khoshgoftaar, T. M., Dittman, D. J., Wald, R., and Awada, W. A review of ensemble classification for DNA microarrays data. Proceedings - International Conference on Tools with Artificial Intelligence, 04-06 November 2013, Herndon, VA, USA, ICTAI. 2013;381–9.

Khoshgoftaar, T. M., Golawala, M., and van Hulse, J. (2007). An empirical study of learning from imbalanced data using random forest. *Proc. - Int. Conf. Tools Artif. Intell. ICTAI.* 2, 310–317. doi:10.1109/ICTAI.2007.49

Ko, L. J., and Prives, C. (1996). p53: Puzzle and paradigm. *Genes Dev.* 10 (9), 1054–1072. doi:10.1101/gad.10.9.1054

Kostic, M., Matt, T., Martinez-Yamout, M. A., Dyson, H. J., and Wright, P. E. (2006). Solution structure of the Hdm2 C2H2C4 RING, a domain critical for ubiquitination of p53. *J. Mol. Biol.* 363 (2), 433–450. doi:10.1016/j.jmb.2006.08.027

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., et al. (2018). ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* 46 (D1), D1062–D1067. doi:10.1093/nar/gkx1153

Levine, A. J., and Oren, M. (2009). The first 30 years of p53: Growing ever more complex. *Nat. Rev. Cancer* 9 (10), 749–758. doi:10.1038/nrc2723

Lim, S., Tan, S. J., Lim, W. T., and Lim, C. T. (2019). Compendiums of cancer transcriptomes for machine learning applications. *Sci. Data* 6 (1), 194–198. Available from. doi:10.1038/s41597-019-0207-2

Liu, J., Sukhova, G. K., Sun, J. S., Xu, W. H., Libby, P., and Shi, G. P. (2004). Lysosomal cysteine proteases in atherosclerosis. *Arterioscler. Thromb. Vasc. Biol.* 24 (8), 1359–1366. doi:10.1161/01.ATV.0000134530.27208.41

Malkin, D., Li, F. P., Strong, L. C., Fraumeni, J. F., Nelson, C. E., Kim, D. H., et al. (1979). Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science* 250 (4985), 1233–1238. doi:10.1126/science.1978757

Mantovani, F., Collavin, L., and del Sal, G. (2019). Mutant p53 as a guardian of the cancer cell. *Cell Death Differ.* 26 (2), 199–212. Available from. doi:10.1038/s41418-018-0246-9

Maurya, N. S., Kushwaha, S., Chawade, A., and Mani, A. (2021). Transcriptome profiling by combined machine learning and statistical R analysis identifies TMEM236 as a potential novel diagnostic biomarker for colorectal cancer. *Sci. Rep.* 11, 14304. doi:10.1038/s41598-021-92692-0

Maßberg, D., and Hatt, H. (2018). Human olfactory receptors: Novel cellular functions outside of the nose. *Physiol. Rev.* 98 (3), 1739–1763. doi:10.1152/physrev.00013.2017

Moyer, S. M., Wasylishen, A. R., Qi, Y., Fowlkes, N., Su, X., and Lozano, G. (2020). P53 drives a transcriptional program that elicits a non-cell-autonomous response and alters cell state *in vivo*. *Proc. Natl. Acad. Sci. U. S. A.* 117 (38), 23663–23673. doi:10.1073/pnas.2008474117

Murray, D., and Mirzayans, R. (2020). Cellular responses to platinum-based anticancer drugs and UVC: Role of P53 and implications for cancer therapy. *Int. J. Mol. Sci.* 21 (16), 5766. doi:10.3390/ijms21165766

NCI (2022). NCI drug dictionary [internet]. *Natl. Cancer Inst.* Available from: https://www.cancer.gov/publications/dictionaries ([cited 2022 Jan 24].

Nikulenkov, F., Spinnler, C., Li, H., Tonelli, C., Shi, Y., Turunen, M., et al. (2012). Insights into p53 transcriptional function via genome-wide chromatin occupancy and gene expression analysis. *Cell Death Differ.* 19 (12), 1992–2002. doi:10.1038/cdd.2012.89

NM_000546 National center for biotechnology information. ClinVar. (p.Thr125=) [Internet][VCV000237948.3]Available from: https://www.ncbi.nlm.nih.gov/clinvar/variation/VCV000237948.3 ([cited 2022 Jan 9].6TP53c.375G>T.

NM_000546. *Natl. Cent. Biotechnol. Inf. ClinVar* 6 (TP53), 672G>A. (p.Glu224=)[VCV000080709.6].):c.

NM_000546. *Natl. Cent. Biotechnol. Inf. ClinVar* 6 (TP53), c.993G>A. (p.Gln331=)[VCV000428868.7].

NM_000546. *Natl. Cent. Biotechnol. Inf. ClinVar* 6 (TP53), 207T>C. (p.Ala69=)[VCV000219841.7].):c.

NM_000546. Thr125=). *Natl. Cent. Biotechnol. Inf. ClinVar* 6 (TP53), c.375G>A.[VCV000177825.18].

NM_000546. Thr125=). *Natl. Cent. Biotechnol. Inf. ClinVar* 6 (TP53), c.375G>C.[VCV000480746.3].

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *J. ofMachine Learn. Res.* 12, 2825–2830. doi:10.5555/1953048.2078195

Piette, J., Neel, H., and Maréchal, V. (1997). Mdm2: Keeping p53 under control. *Oncogene* 15 (9), 1001–1010. doi:10.1038/sj.onc.1201432

Pleasance, E., Titmuss, E., Williamson, L., Kwan, H., Culibrk, L., Zhao, E. Y., et al. (2020). Pan-cancer analysis of advanced patient tumors reveals interactions between therapy and genomic landscapes. *Nat. Cancer* 1 (4), 452–468. doi:10.1038/s43018-020-0050-6

Prives, C., and Hall, P. A. (1999). The P53 pathway. *J. Pathol.* 187 (1), 112–126. doi:10.1002/(SICI)1096-9896(199901)187:1<112:AID-PATH250>3.0.CO;2-3

Reel, P. S., Reel, S., Pearson, E., Trucco, E., and Jefferson, E. (2021). Using machine learning approaches for multi-omics data analysis: A review. *Biotechnol. Adv.* 49, 107739. doi:10.1016/j.biotechadv.2021.107739

Riley, T., Sontag, E., Chen, P., and Levine, A. (2008). Transcriptional control of human p53-regulated genes. *Nat. Rev. Mol. Cell Biol.* 9 (5), 402–412. doi:10.1038/nrm2395

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., et al. (2011). Integrative genomics viewer. *Nat. Biotechnol.* 29 (1), 24–26. doi:10.1038/nbt.1754

Saghaleyni, R., Muhammad, A. S., Bangalore, P., Nielsen, J., and Robinson, J. L. (2021). Machine learning-based investigation of the cancer protein secretory pathway. *PLoS Comput. Biol.* 17 (4), e1008898–20. Available from. doi:10.1371/journal.pcbi.1008898

Santamaría, I., Velasco, G., Pendás, A. M., Fueyo, A., and López-Otín, C. (1998). Cathepsin Z, a novel human cysteine proteinase with a short propeptide domain and a unique chromosomal location. *J. Biol. Chem.* 273 (27), 16816–16823. doi:10.1074/jbc.273.27.16816

Seong, H. A., and Ha, H. (2012). Murine protein serine-threonine kinase 38 activates p53 function through Ser 15 phosphorylation. *J. Biol. Chem.* 287 (25), 20797–20810. doi:10.1074/jbc.M112.347757

Shu, C., Zheng, X., Wuhafu, A., Cicka, D., Doyle, S., Niu, Q., et al. (2022). Acquisition of taxane resistance by p53 inactivation in ovarian cancer cells. *Acta Pharmacol. Sin.* doi:10.1038/s41401-021-00847-6

Shvarts, A., Steegenga, W. T., Riteco, N., van Laar, T., Dekker, P., Bazuine, M., et al. (1996). Mdmx: A novel p53-binding protein with some functional properties of MDM2. *EMBO J.* 15 (19), 5349–5357. doi:10.1002/j.1460-2075.1996.tb00919.x

Smith, A. M., Walsh, J. R., Long, J., Davis, C. B., Henstock, P., Hodge, M. R., et al. (2020). Standard machine learning approaches outperform deep representation learning on phenotype prediction from transcriptomics data. *BMC Bioinforma.* 21 (1), 119–218. doi:10.1186/s12859-020-3427-8

Solyakov, L., Sayan, E., Riley, J., Pointon, A., and Tobin, A. B. (2009). Regulation of p53 expression, phosphorylation and subcellular localization by a G-protein-coupled receptor. *Oncogene* 28 (41), 3619–3630. doi:10.1038/onc.2009.225

Srivastava, S., Zou, Z., Pirollo, K., Blattnert, W., and Chang, E. H. (1990). Germ-line transmission of a mutated p53 gene in a cancer-prone family with Li–Fraumeni syndrome. *Nature* 348, 747–749. doi:10.1038/348747a0

Sun, T., Jiang, D., Zhang, L., Su, Q., Mao, W., and Jiang, C. (2016). Expression profile of cathepsins indicates the potential of cathepsins B and D as prognostic factors in breast cancer patients. *Oncol. Lett.* 11 (1), 575–583. doi:10.3892/ol.2015.3960

Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T., and Lehner, B. (2014). Synonymous mutations frequently act as driver mutations in human cancers. *Cell* 156 (6), 1324–1335. doi:10.1016/j.cell.2014.01.051

Tanikawa, C., Ri, C., Kumar, V., Nakamura, Y., and Matsuda, K. (2010). Crosstalk of EDA-A2/XEDAR in the p53 signaling pathway. *Mol. Cancer Res.* 8 (6), 855–863. doi:10.1158/1541-7786.MCR-09-0484

Thomas, A. F., Kelly, G. L., and Strasser, A. (2022). Of the many cellular responses activated by TP53, which ones are critical for tumour suppression? *Cell Death Differ.* 29 (5), 961–971. doi:10.1038/s41418-022-00996-z

Vardanyan, R., and Hruby, V. (2016). "Antineoplastic Agents," in *Synthesis of best-seller drugs*. Editors R. Vardanyan and V. Hruby (Cambridge: Academic Press), 495–547.

Vassilev, L. T. (2007). MDM2 inhibitors for cancer therapy. *Trends Mol. Med.* 13 (1), 23–31. doi:10.1016/j.molmed.2006.11.002

Vousden, K. H., and Lu, X. (2002). Live or let die: The cell's response to p53. *Nat. Rev. Cancer* 2 (8), 594–604. doi:10.1038/nrc864

Vousden, K. H., and Prives, C. (2009). Blinded by the light: The growing complexity of p53. *Cell* 137 (3), 413–431. doi:10.1016/j.cell.2009.04.037

Waskom, M. (2021). Seaborn: Statistical data visualization. *J. Open Source Softw.* 6 (60), 3021. doi:10.21105/joss.03021

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* 45 (10), 1113–1120. doi:10.1038/ng.2764

Wilke, R. A., and Dolan, M. E. (2011). Genetics and variable drug response. *JAMA - J. Am. Med. Assoc.* 306 (3), 306–307. doi:10.1001/jama.2011.998

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2018). DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res.* 46 (D1), D1074–D1082. doi:10.1093/nar/gkx1037

Wu, Y., and Wang, G. (2018). Machine learning based toxicity prediction: From chemical structural description to transcriptome analysis. *Int. J. Mol. Sci.* 19 (8), E2358. doi:10.3390/ijms19082358

Xiong, X., Zhao, Y., He, H., and Sun, Y. (2011). Ribosomal protein S27-like and S27 interplay with p53-MDM2 axis as a target, a substrate and a regulator. *Oncogene* 30 (15), 1798–1811. doi:10.1038/onc.2010.569

Yuan, B., Yang, D., Rothberg, B. E. G., Chang, H., and Xu, T. (2020). Unsupervised and supervised learning with neural network for human transcriptome analysis and cancer diagnosis. *Sci. Rep.* 10 (1), 19106–19111. Available from. doi:10.1038/s41598-020-75715-0

Zhang, A., Liu, C., and Lin, G. (2021). P53 pathway activate detection based on machine learning: The modified XGBoost-based method of pan-cancer pathway activity detection in the cancer genome atlas. *CCEAI* 2021, 41–45. doi:10.1145/3448218.3448237

Zhang, L., Luo, M., Yang, H., Zhu, S., Cheng, X., and Qing, C. (2019). Next-generation sequencing-based genomic profiling analysis reveals novel mutations for clinical diagnosis in Chinese primary epithelial ovarian cancer patients. *J. Ovarian Res.* 12 (1), 19–9. doi:10.1186/s13048-019-0494-4

Zheng, L. Z., and Chen, S. Z. (2018). shRNA-induced knockdown of the SPERT gene inhibits proliferation and promotes apoptosis of human colorectal cancer RKO cells. *Oncol. Rep.* 40 (2), 813–822. doi:10.3892/or.2018.6455