HHS Public Access

Author manuscript

Cell Rep. Author manuscript; available in PMC 2020 July 23.

Published in final edited form as:

Cell Rep. 2019 January 08; 26(2): 496–506.e3. doi:10.1016/j.celrep.2018.12.066.

Characterization of Tumor-Suppressor Gene Inactivation Events in 33 Cancer Types

Peilin Jia[1], **Zhongming Zhao**[1,2,3,4,*]

[1]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

[2]MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences, Houston, TX 77030, USA

[3]Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

[4]Lead Contact

SUMMARY

We systematically investigated the landscape of tumor-suppressor gene (TSG) inactivation events in 33 cancer types by quantitatively measuring their global and local genomic features and their transcriptional and signaling footprints. Using The Cancer Genome Atlas data, we identified with high confidence 337 TSG × cancer events in 30 cancer types, of which 277 were unique events. The majority (91.0%) of these events had a significant downstream impact measured by reduced expression of the TSG itself (cis-effect), disturbance of the transcriptome (trans-effect), or combinatorial effects. Importantly, the transcriptomic changes associated with TSG inactivation events were stronger than the cancer lineage difference, and the same TSGs inactivated in different cancer types tended to cluster together. Several TSGs (e.g., *RB1*, *TP53*, and *CDKN2A*) involved in the regulation of the cell-cycle-formed clusters. Finally, we constructed subnetworks of the TSG × cancer inactivation events, including the local genes frequently disturbed upon the inactivation events.

INTRODUCTION

Cancer development involves multiple dysregulated processes leading to uncontrolled cell growth (Vogelstein et al., 2013). Activation of oncogenes and inactivation of tumor-suppressor genes (TSGs) are two major driving forces in cancer (Bowden et al., 1994). To

*Correspondence: zhongming.zhao@uth.tmc.edu.

AUTHOR CONTRIBUTIONS

P.J. and Z.Z. conceived of the project. P.J. performed the data analyses. P.J. and Z.Z. wrote the manuscript.

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures and can be found with this article online at https://doi.org/10.1016/j.celrep.2018.12.066.

DECLARATION OF INTERESTS

The authors declare no competing interests.

date, more than a thousand human genes have been reported to be TSGs or to play tumor-suppressor roles (Davoli et al., 2013; Zhao et al., 2013, 2016). Recent high-throughput projects, such as The Cancer Genome Atlas (TCGA), have led to the discovery of many novel TSGs (Davoli et al., 2013). The functions of TSGs are broadly distributed across various cellular processes, including signaling pathways, chromatin remodeling, and DNA damage and repair processes, among others (Zhao et al., 2016). However, a systematic investigation of the molecular profiles and functional impacts of TSG inactivation events is still lacking. Presently, TSGs are much more difficult than oncogenes to approach as drug targets. Therefore, understanding the downstream impacts of TSGs will likely indicate new directions of drug targets (Cheng et al., 2016).

The two-hit model has been commonly applied to study the roles of TSGs in cancer (Knudson, 1971). Unlike oncogenes, most loss-of-function (LoF) mutations that occur in TSGs act in a recessive way in nature. In principle, both copies of a TSG need to be mutated for a TSG to completely lose its function, whereas haploinsufficient TSGs work in a dosage-sensitive manner. Mutations that may lead to LoF of their residing genes include, but are not limited to, nonsense single nucleotide variants (SNVs), SNVs disrupting canonical splicing sites, small insertions and deletions (indels) that cause frameshift, and copy-number variations (CNVs) (Johnston et al., 2015). In cells, mRNA carrying a nonsense mutation undergoes nonsense-mediated decay (NMD) (Schweingruber et al., 2013), resulting in a decrease of its expression; however, some genes may escape NMD. Other mechanisms, such as epigenomic regulation (e.g., hyper-methylation of its promoter region) and post-transcriptional regulation (e.g., microRNA suppression), may lead to an inactivation of a TSG. For the haploinsufficient mechanism in TSGs, one copy of the gene is inactivated, and the remaining one is inadequate to maintain the normal function of the gene (Davoli et al., 2013).

In this study, we performed a systematic, integrative analysis of TSG inactivation events and further assessed their impact on transcription and signaling pathways. We proposed a computational framework to quantitatively measure genetic events that lead to TSG inactivation. We examined their impact on the whole transcriptome, the disturbed signaling pathways, and the local subnetworks. Our study provides a landscape view of TSG alterations and their potential impact in 30 cancer types.

RESULTS

Definition of a TSG Inactivation Event

We proposed a framework to define genetic inactivation mutations in TSGs using both CNVs and SNVs/indels (Figure 1). We defined two types of inactivation mutations with evidence that is consistent with the two-hit model. The first type is based on the evidence in supporting the loss of both copies due to genetic mutations (hereafter, we called L2). L2 mutations included two subtypes: (a) a deep deletion (CN = −2, CN: copy number) and (b) a shallow deletion (CN = −1), which simultaneously occurs with at least one truncation mutation (nonsense SNV, splice site SNV, or frameshift indel). The second type includes the cases in which one copy of the gene has solid evidence of loss, whereas the other copy has less compelling, but still supportive, evidence for LoF (called L1). L1 mutations include

three subtypes: (a) a shallow deletion (CN = −1) accompanied with a deleterious missense SNV, (b) a shallow deletion (CN = −1) accompanied with an in-frame indel, and (c) a truncation mutation in non-CN-loss samples (CN ≥ 0). This framework can be further refined when additional data become available. For example, mutations categorized as L1 (c) could lead to the loss of both copies if they were homozygous truncation mutations but such information requires raw sequencing data and accurate variant calling procedures. In this study, we considered these mutations as inadequate evidence and categorized them as L1. For each TSG, we identified and labeled tumor samples by L1 and L2 mutations, respectively. We also identified its wild-type (WT) samples as those that did not undergo copy-number loss (CN ≥ 0) and did not have any truncation mutations. For WT samples, they were allowed to harbor benign missense SNVs or in-frame indels, because such mutations would not lead to simultaneous loss of two copies of the TSG. The remaining samples included those with only copy-number loss but no mutations or those with deleterious missense SNVs but no copy-number loss. For TSGs on chromosome X, the shallow deletion was applied to female samples only.

We collected 1,229 protein-coding TSGs, including 40 on chromosome X. We processed the TCGA data for 33 cancer types (full names of cancer types are provided in Figure 2). Particularly, BRCA samples were separated into five subtypes due to strong subtype-driven transcriptome impact and its available large sample size. Four cancer types were excluded due to small sample size (Figure S1). Therefore, we analyzed data in 33 cancer types/subtypes (30 cancer types). Specifically, 1,020 (83.0%, 1,020/1,229) TSGs had L2 or L1 genetic inactivation events in at least five samples from the same cancer type among a total of 5,510 samples (76.0% of 7,248 samples with qualified data; Figures 2A and S2). Throughout this work, only the TSGs that were inactivated in at least five samples in each cancer type were investigated. The 40 most frequently inactivated TSGs are shown in Figure 2A. Several clusters were observed, such as 8p21–23, 9p (containing *CDKN2A/CDKN2B*), and 13q (containing *RB1*), consistent with previous reports (Ciriello et al., 2013). Although several TSGs located in these regions could be driver genes, such as *SOX7* (Man et al., 2015), *NKX3–1* (Bowen et al., 2013), *CLU* (Chayka et al., 2009), and *TRIM35* (Chen et al., 2015), many were likely passenger events because of large CNV segments. Accordingly, statistical analysis strategies are required to assess and prioritize candidate driver TSGs.

Genetic Inactivation Events

We identified three groups of genomic factors that may affect the determination of TSG inactivation events: sample-based bias, gene-length bias, and putative inflation sites in the genome. Sample-based bias addresses the cases in which genes had inactivation mutations predominantly in the samples with a high mutation load. A weighted resampling strategy was applied and a p_{sample} value for each TSG in each cancer type was calculated, wherever applicable (i.e., sample size ≥ 20). TSGs with $p_{\text{sample}} < 0.05$ for L2 or L1, respectively, were selected as candidates. In this adjustment, the mutations in those samples were unlikely due to random mutation accumulation in their tumor genomes. The gene-length has been considered as a bias in many genome analyses, for example, the longer genes tend to have more mutations to be detected from genome sequence data. We fitted a smooth model in each cancer type, where the gene frequency was regressed on its length, and calculated

p_{gene}. We chose TSGs with $p_{gene} < 0.05$, indicating that the mutation frequency in these TSGs remained significant after adjusting for gene length. Finally, we identified the loci that were frequently reported as CNV loss and had manual examination. For example, we removed five genes close to telomeres. These genes were reported with deep deletions, but our manual check did not find the short reads mapped to those gene regions (Figure S3). Furthermore, 22 extremely long TSGs (> 800 kb), including six located near the chromatin fragile sites (*CNTNAP2*, *CTNNA3*, *DMD*, *FHIT*, *PARK2*, and *WWOX*), were also removed. Through these processes, we identified 337 significant TSG inactivation events in 30 cancer types (Figure 2D), involving 138 unique TSGs ($p_{sample} < 0.05$ and $p_{gene} < 0.05$). Hereafter, a TSG inactivated in a specific cancer type is referred to as a TSG × cancer inactivation event.

The most frequently inactivated TSGs included *TP53* (20 cancer types), *CDKN2A* (18) and its colocalization partners *CDKN2B* (17) and *MTAP* (15), *PTEN* (15), and *RB1* (14). The features of inactivation mutations varied among TSGs (Figure 2D). For example, *CDKN2A*, *CDKN2B*, and *MTAP* were located together on chromosome 9 and were predominantly inactivated by homozygous CNV loss (see oncoprints in Figures S4 and S5). The co-deletion of these three genes has been previously reported (Mavrakis et al., 2016). Most other genes were not observed with such a consistent mutation type, but rather inactivated by various kinds of mutations as shown with three or more colorful stacks in Figure 2D. Although deleterious missense mutations were included, they were found mainly in a few genes, such as *TP53* (multiple cancer types), *PTEN* (GBM), *KEAP1* (LUAD), and *VHL* (KIRC). Most genes were found with both L2 and L1 mutations, as shown with four colorful stacks in Figure 2D. Many genes with deep deletion were filtered out by the sample-based test because they mainly, and sometimes only, occurred in samples with a high mutation load. Hence, they were likely passengers resulting from genome-wide accumulation of mutations.

To assess the biological significance of TSG × cancer inactivation events, we compared the *cis*- and the *trans*-effect between significant and non-significant TSG × cancer events using TCGA transcriptome data (Figure 3A). Sixty cases had sufficient data (≥5 testable TSGs and ≥ 3 significant TSGs). Among them, the significant TSGs in 48 cases had a stronger *cis*-effect and a *trans*-effect than did non-significant TSGs (Figure S6). This indicated that the TSG × cancer inactivation events identified by our methods at the DNA level could be supported by their gene expression.

When two or more L2 TSGs are located in a region overlapped with a CNV deletion, we clustered these TSGs and selected one to represent each cluster. For example, in the *CDKN2A*-*CDKN2B*-*MTAP* cluster, we selected *CDKN2A* for the following analysis. In this process, we obtained 277 TSG × cancer events for transcriptomic analyses, representing unique inactivation events.

TSG Inactivation Events at the Transcription Level

Next, we explored the impact of genetic inactivation events on the TSGs themselves (*cis*-effect) and on other genes (*trans*-effect) using the transcriptome data. For the *cis*-effect, although we expected that nonsense SNVs and copy-number loss would result in decreased expression, this was not always the case in the actual data because of various reasons, such

as unusual biological mechanisms (e.g., escaping NMD or stop codon read through) (Holbrook et al., 2004), technique issues (e.g., sequencing errors and sample purity), and variant calling accuracy. In this work, a *trans*-effect is defined as a TSG × cancer event that influences the expression of other genes, such as downstream genes in the signaling pathways. It is commonly recognized that the passenger mutations irrelevant to cancer development would not have much impact on other genes, whereas driver mutations are expected to interrupt signaling pathways that are critical to cancer. To measure both *cis*- and *trans*-effects, a differential expression analysis for each TSG × cancer event was performed by comparing the expression profiles of L2+L1 samples with the matched WT samples for each TSG, resulting in transcriptome-wide *t*-values for all these genes. We defined a combined impact score $T = \sqrt{t_1^2 + t_2^2}$, where t_1 is the *t*-value for the TSG itself (*cis*-effect) and t_2 is the average *t*-value for the top 1% differentially expressed genes (DEGs) in the transcriptome excluding the TSG itself (*trans*-effect).

Further filtering was performed to remove abnormal outliers. As a result, 208/277 (75.1%) events remained unchanged, 53 (19.1%) were reduced by less than 4 samples, 5 (1.8%) were reduced by 4–7 samples, and 11 events were excluded due to insufficient number of samples (Figure S7). Hereafter, we use inactivated samples to refer to the samples from L2+L1 after this filtering process.

To explore the distribution of *cis*- and *trans*-effects, we separated the remaining 266 TSG × cancer events into four subgroups according to whether their *cis*-effect was nominally significant ($p_{cis} < 0.05$, corresponding to t_1 in DEG analysis) and whether their *trans*-effect was significant ($z_{trans} > 1.28$, corresponding to FDR < 0.20, where z_{trans} was calculated as the average *z*-score transferred from p_{BH} from DEG analysis) (Figure 3D). First, the most common pattern included TSG × cancer events with both a *cis*-effect and a *trans*-effect (n = 150). This finding is expected and supports that genetic inactivation events lead to decreased expression of TSGs (*cis*-effect), which in turn regulates their downstream genes (*trans*-effect). Mutations in these events are more likely to contribute to cancer. Second, 39 TSG × cancer events had no *cis*-effect but had a substantial *trans*-effect (blue dots in Figure 3D). We illustrate this using two cases: *MAP3K1* (Figure 3B) and *GATA3* (Figure 3C). *MAP3K1* had inactivation mutations in 44 BRCA_LumA samples but no sign of decreased expression: all with truncation mutations (L1). However, broad DEGs were observed in these 44 samples compared with *MAP3K1* WT samples. A similar trend was observed for *GATA3* in both BRCA_LumA (51 inactivated samples with truncation mutations (L1), not shown in the figure) and BRCA_LumB (29 inactivated samples with truncation mutations, L1). Third, 53 TSG × cancer events had a *cis*-effect, but not much of a *trans*-effect (green dots in Figure 3D). Examples included *KDM6A* in BLCA, ESCA, HNSC, LUSC, and PAAD; *ARID1A* in BLCA, LIHC, and LUAD; and *NCOR1* in LIHC and STAD. Finally, there were 24 TSG × cancer events with neither a *cis*-effect nor a *trans*-effect (gray dots in Figure 3D). The specific molecular mechanisms of these events on the transcriptome level require further investigation, with possible roles, such as abnormal transcription or splicing. In summary, our analysis revealed that 91.0% TSG × cancer events had effect in the forms of a *cis*-effect, a *trans*-effect, or both.

TSG Inactivation Events: Global versus Focal Impact

The number of DEGs (#DEGs) for each TSG × cancer event, which represents a direct measurement of the TSG's *trans*-effect, varied dramatically (Figure 3E). We defined DEGs as those with fold change (FC) > 2 and $p_{BH} < 0.2$ (p_{BH}: FDR values by Benjamini-Hochberg (BH) multiple test correction). In 29 TSG × cancer events, we observed a global impact with more than 500 DEGs. Examples included *RB1* in LUAD ($n_{mut} = 28$, #DEGs = 627), *VHL* in KIRC ($n_{mut} = 110$, #DEGs = 3,765), *PBRM1* in KIRC ($n_{mut} = 98$, #DEGs = 2,865), *SETD2* in KIRC ($n_{mut} = 40$, #DEGs = 2,409), *BAP1* in KIRC ($n_{mut} = 39$, #DEGs = 2,339), and *SMAD4* in PAAD ($n_{mut} = 51$, #DEGs = 1,154). Overall, 143 TSG × cancer events broadly affect a scale of 20–500 DEGs, such as *RB1*, *NF1*, *BRCA1*, and *BRCA2*. The remaining 94 TSG × cancer events had less than or equal to 20 DEGs, implying a weak *trans*-effect. The variation in DEGs was unlikely due to the small number of inactivated samples, because a similar trend of DEGs was still observed when requiring $n_{mut} \geq 20$ for each TSG × cancer event (Figure 3G). These results suggested that the strong variation of DEGs was likely due to TSG's biological disturbance rather than random sampling.

To evaluate the significance of the observed transcriptomic footprints of these TSG × cancer events, we implemented a randomization test and found that in 161 (60.5%) TSG × cancer events (71 unique TSGs), the TSGs had significantly stronger transcriptomic impact than what would be randomly expected (#DEGs ≥ 10, $p_{emp} < 0.05$) (Figure 3F).

Strong TSG Inactivation Characteristics beyond Cancer Lineages

Next, we explored whether inactivation of the same TSG would lead to similar transcriptional footprints in different cancer types. We used transcriptome-wide *t*-values of each TSG × cancer event to perform a hierarchical cluster analysis and particularly focused on the 161 TSG × cancer events with a significant impact (#DEGs ≥ 10, $p_{emp} < 0.05$). An initial application illustrated two distinctive groups, where *RB1*, *TP53*, and *CDKN2A* were frequently observed to cluster together but within-cancer groups were also observed (Figure S7). To control for potential confounding factors of shared samples (e.g., one sample with inactivated events by two TSGs), we restricted the analyses only to those samples that did not share mutant TSGs. Specifically, when comparing two TSGs, e.g., TSG_1 (inactivated samples denoted by MT_1 and wild-type samples by WT_1) and TSG_2 (MT_2 and WT_2), we chose to use only MT_1 and WT_1 in the wild-type samples of TSG_2 (WT_2), and vice versa. As a result, for any pair of TSGs compared, the transcriptome-wide *t*-values for a TSG were only obtained in the wild-type of the other TSG, and, hence, the potential confounding impact by the other TSG was controlled. Using the transcriptome-wide *t*-values for each TSG × cancer event, we conducted hierarchical clustering and observed two major clusters (Figure 4A). TSGs involved in cell-cycle regulation, such as *TP53*, *RB1*, *CDKN2A*, and *PTEN*, were overrepresented in the larger group. Although some cancer types were occasionally enriched in local subclusters (e.g., LGG and KIRC), the overall cluster showed enrichment independent of cancer types, suggesting common pathways might be disturbed upon the inactivation of the same TSGs in different cancer types. In addition to the *RB1*, *PTEN*, *CDKN2A*, and *TP53* clusters, we also observed local structures, such as *STK11*, *FBXW7*, *BAP1*, and *NCOR1*.

Cell-Cycle Pathway

The results above included several TSGs in the cell-cycle pathway. Here, we specifically examined the expression profiles of the genes in this pathway. It has been reported that *RB1* status is inversely correlated with *CDKN2A* (Shapiro et al., 1995), yet changes in the expression levels of both *RB1* and *CDKN2A* could lead to disturbance of the cell-cycle pathway. We questioned whether the component genes of this pathway showed any expression relationship when *RB1* was inactivated and when *CDKN2A* was inactivated. To control potential confounding effects, we restricted *RB1* mutant and WT samples both in *CDKN2A* WT samples and vice versa. As shown in Figure 4B, we observed a consistent positive correlation of the cell-cycle genes (excluding *RB1* and *CDKN2A/CDKN2B* themselves) in all cancer types, except for HNSC. *CDKN2A* (and its neighbor gene *CDKN2B*) showed an inverse relationship with *RB1*, which is consistent with previous reports (Møller et al., 1999). However, all other genes were positively correlated, indicating that inactivating either *RB1* or *CDKN2A* had a similar downstream impact on the cell-cycle pathway.

Mizuarai et al. (2011) reported that the ratio between *CDKN2A* and *CCND1* is predictive of *RB1* status. A higher *CDKN2A/CCND1* ratio implies that *RB1* is either absent or inactivated. We examined the expression difference of these two genes (*CDKN2A* and *CCND1*) and confirmed its prediction feature: the *RB1*-inactivated samples had much higher expression difference between *CDKN2A* and *CCND1* than *RB1* WT samples. Furthermore, in *RB1*-inactivated samples, *BRCA1* was upregulated. That *BRCA1* is a critical gene in the DNA repair pathway may imply that the DNA repair pathway was also activated in *RB1*-inactivated samples.

Signaling Pathways Disturbed by TSG Inactivation Events

To examine significantly disturbed pathways, we performed a single sample gene set enrichment analysis (ssGSEA) (Hänzelmann et al., 2013) for each TSG × cancer event. We used 1,044 canonical pathways from BioCarta, Reactome, KEGG, and the Pathway Interaction Database (Subramanian et al., 2005). ssGSEA calculates an enrichment score (ES) for each pathway in each sample. We used Wilcoxon rank-sum test to compare the ES in TSG inactivated samples and the ES in WT samples for each TSG for each pathway. Figure 4E displays significant pathways (Bonferroni corrected p value < 0.05) enriched in more than 16 TSG × cancer events (10% of all 161 events). Two groups of pathways were highlighted. One brings together cell cycle and its related processes, and the other for immune related pathways. Cell-cycle and related pathways were mainly disturbed in canonical TSGs, such as *RB1*, *TP53*, and *PTEN*, including cell-cycle checkpoints, DNA replication, unwinding of DNA, synthesis of DNA, mismatch repair, MCM pathway, and E2F pathway. The immune-related pathway group included the interleukin-5 (IL-5) pathway, the IL-7 pathway, the IL-12 pathway, the TH1TH2 pathway, PD1 signaling, the cytotoxic T lymphocyte (CTL) pathway, and the TCRA pathway. However, these pathways were predominantly enriched in a few events, including *APC* (COAD), *BAP1* (KIRC), *CDKN2A* (SKCM), *FUBP1* (LGG), *NF1* (LGG), *RGS12* (LUSC), *SETD2* (KIRC), *SMAD4* (STAD), and *TP53* (HNSC).

Disturbed Local Networks of TSG Events Support TSG Cluster Features

By imposing *t*-values onto a protein-protein interaction (PPI) network, we determined the disturbed partners to each TSG × cancer event. We took *RB1* as an example, which had significant impact in 14 cancer types. We obtained its partners in the network that were significantly disturbed in each cancer type ($z_i^B \geq 2$ and $z_i^\pi \geq 2$; see the STAR Methods), resulting in 6 (LUSC) to 141 genes (LIHC). Interestingly, 50 of these genes were recurrently disturbed in at least five cancer types (Figure 5A). These genes formed two major groups in the subnetwork: one from the cell-cycle pathway (e.g., *CCNB1*, *CCND1*, *CCNE1*, *CCNE2*, *CDK1*, *CDK2*, *E2F1*, *MCM2*, and *MCM3*) and other from the DNA damage repair pathway (e.g., *MSH2*, *MSH6*, *FANCA*, *POLE*, *POLA1*, and *RAD51*). Similarly, the affected subnetwork for *TP53*-highlighted genes in the cell-cycle pathway and the *TP53* signaling pathway (Figure 5B).

TSG Inactivation Associated with Deleterious Missense Mutations

According to our definition of L2 and L1, deleterious missense mutations were not included unless they co-occurred with a CNV loss (CN = −1 or −2). However, some TSGs had a large number of deleterious missense mutations. When these mutations occurred in copy neutral samples, they would not be classified as either inactivation (L2 or L1) or WT samples (hence, excluded in transcriptional analysis). To explore the impact of deleterious missense mutations, we specifically examined 56 such TSG × cancer cases involving 38 unique TSGs. For these cases, at least five samples had deleterious missense mutations without a copy loss. We tested whether these mutations could also lead to inactivation events of TSGs. To this end, we performed a DEG analysis for the following: (1) the inactivated samples versus WT samples and (2) the samples having only missense mutations versus WT samples. The analysis revealed 21 TSG × cancer cases involving 9 TSGs had a Jaccard Index score > 0.05 based on shared DEGs. This result indicated that deleterious missense mutations in these genes and the inactivation mutations led to a shared *trans*-effect, and the effect was higher than expected by chance. These TSGs included *ATRX* (LGG), *BAP1* (UVM), *CASP8* (HNSC), *CDKN2A* (HNSC), *EP300* (BLCA), *KEAP1* (LUAD), *RUNX1* (LAML), *STK11* (LUAD), and *TP53* (multiple cancer types). Figure 6 shows three example genes, *TP53*, *ATRX*, and *KEAP1*. Interestingly, even though those deleterious missense mutations had some *trans*-effect, they did not display a *cis*-effect. Taken together, these results suggested that the deleterious mutations in these genes also might lead to inactivation of the corresponding TSGs. Therefore, this feature should be considered in searching TSG inactivation events from future cancer genomic data.

DISCUSSION

We performed a systematic analysis and characterized the functional features of TSG inactivation events in more than 5,000 tumor genomes of 33 cancer types/subtypes. We described the spectrum of genetic alterations leading to TSG inactivation; classified a *cis*-effect and a *trans*-effect for each TSG event; and investigated the potential impact of TSG events on the transcriptome, in signaling pathways, and in protein interaction networks. Consistent expression patterns of TSGs were observed across cancer types, and TSG group features overwhelmed the similarity of the cancer lineages (Figure 4A). In addition to the

well-studied TSGs, such as *RB1* and *TP53*, our results pinpointed TSGs functioning in various processes of epigenomic regulation (e.g., *ARID1A*, *ARID1B*, and *KDM6A*). These included chromatin remodeling genes, such as *ARID1A*, *SETD2*, *KDM5C*, and *KDM6A*, and genes involved in transcription, such as *RUNX1* and *GATA3*. These results have critical implications for identification and interpretation of TSG inactivation events in cancer.

There are a few limitations in this study. First, some mutation types could not be included in our framework. For example, loss of heterozygosity (LOH) is a common event associated with inactivation of TSGs in many cancer types (Ryland et al., 2015). LOH includes copy-number loss LOH and copy number neutral LOH. In this study, we did not explicitly model LOH events, because we did not have access to the raw data. The model presented could partially integrate LOH events when a somatic inactivation mutation was detected (either a nonsense SNV or a frameshift indel). In such cases, the mutation would be included in the model as either WT_NS_FS mutation type in L1 (it may be considered equivalent to copy-neutral LOH) or Hetlose_NS_FS mutation type in L2 (it may include copy-loss LOH). However, if a mutation was inherited and was not in the somatic mutation calls, we were unable to include it in the model. In addition, if one sample contained multiple NS_FS mutations in one TSG, the sample was considered as inactivated only once and would not be assigned a stronger weight, which might underestimate the severity of the inactivation mutations in some cases. Second, the sample-based test or gene-based test likely excluded some candidate TSGs. Some TSGs also function in DNA damage repair (DDR) deficiency, such as *BRCA2*, *ATM*, and *TP53*. Inactivation of these genes may result in increased genomic instability and increased mutation load. The sample-based test matches the mutation load in randomization sets and hence may falsely reject such TSG/DDR genes. This study focused on transcriptomic impact. A more comprehensive analysis of DDR genes is described in Knijnenburg et al. (2018).

Third, we only considered genetic events. Other mechanisms, such as epigenomic inactivation through methylation, post-transcriptional regulation, could lead to inactivation of TSGs. In addition, some TSGs also function as oncogenes in certain cases, such as *TP53* and *PTEN*. Distinguishing which role the TSG plays and whether a mutation leads to activation or inactivation of the TSG remains a challenge for future work. Finally, the task to distinguish the impacts of TSG inactivation event is complicated as shown in the analysis. We examined the transcriptional footprint in the forms of a *cis*- and a *trans*-effect, global versus local impacts, lineage versus TSG similarities, and downstream pathways and networks. Future work should investigate other forms of potential impacts, such as mutation load and epigenomic profile.

Despite these limitations, this work provides an important foundation for future TSG investigation. First, we proposed a framework to define TSG inactivation events by combining somatic mutations and CNVs. This framework can be applied to studies with similar designs and can be extended to integrate germline mutations and epigenomic variants (e.g., methylation on promoters). Second, although more than 1,000 TSGs were reported, many of them did not show a significant inactivation event or had a measurable impact on the transcriptome. While several well-known TSGs (e.g., *TP53*, *RB1*, *CDKN2A/CDKN2B*, and *PTEN*) were prevalently inactivated in multiple cancer types, many others were only

inactivated in few cancer types (e.g., *APC* in BLCA, COAD, and READ). Third, our analytical strategy could distinguish *cis*- from *trans*-effect of a TSG inactivation event. A number of TSGs, although genetically inactivated, showed no effect or only one way of effect on the transcriptome. As above, to further explore the features of TSG inactivation events, it is better to examine not only the genetic events but also their functional impacts using transcriptomic, proteomic, and functional genomic data.

In summary, we presented a comprehensive framework to classify TSG inactivation events, revealed the landscape of these events, and extensively explored the potential functional impacts of TSG inactivation events in cancer.

STAR★METHODS

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for data may be directed to and will be fulfilled by the Lead Contact, Zhongming Zhao (zhongming.zhao@uth.tmc.edu).

METHOD DETAILS

Curation of tumor suppressor genes—We compiled a list of 1229 putative TSGs from two sources: (1) the Tumor Suppressor Gene (TSGene) database (version 2.0, n = 1217, retrieved on March 1, 2016) (Zhao et al., 2016), which curates TSGs from NCBI PubMed through manual collection and literature mining; and (2) the list (n = 301) predicted by the computational algorithm TUSON in Davoli et al. (2013). MicroRNA genes and non-coding RNAs were excluded due to limited annotation of somatic mutation data. The two sources had 82 TSGs in common. Some genes function as both tumor suppressor and oncogenes. However, our definitions of L2 and L1 are designed for mutations leading to LoF and all analyses were framed as comparing samples with inactivated mutations versus WT samples. Thus, mutations that may lead to oncogene activation were not included in the analysis.

Multi-dimensional data—We included the TCGA samples that had somatic mutations, CNVs, and mRNA expression data. All data were downloaded from the UCSC Cancer Genome Browser Xena. Figure S1 shows the available samples in each cancer type. The cancer type abbreviations are listed in Figure 2.

For SNVs and indels, we re-annotated the downloaded data using wANNOVAR (Chang and Wang, 2012) Genes whose ratio of nonsynonymous SNVs versus synonymous SNVs (NS/S) was less than 1 were removed in the corresponding cancer type, as NS/S < 1 indicates no positive selection. The longest transcript was used for each protein-coding gene. The results from wANNOVAR provided twelve ways to predict the impact of missense mutations, including SIFT (Kumar et al., 2009), PolyPhen2_HDIV (Adzhubei et al., 2010), PolyPhen2_HVAR (Adzhubei et al., 2010), LRT (Chun and Fay, 2009), MutationTaster (Schwarz et al., 2010), Mutation Assessor (Reva et al., 2011), FATHMM (Shihab et al., 2013), RadialSVM (Dong et al., 2015), LR, GERP (Davydov et al., 2010), phyloP46way (Cooper et al., 2005), and SiPhy_29way_logOdds (Garber et al., 2009). We defined deleterious missense SNVs as those voted by at least two methods. For nonsense SNVs, we

excluded those that may escape NMD by the following three rules: (1) located within the last 50bp of the second to last exon and the whole last exon; (2) located within 200bp downstream of TSS; and (3) located in transcripts with ≤ 2 exons (Hu et al., 2017). Such NMD-escaping mutations may still lead to incomplete proteins, because they would escape degradation at the mRNA level. However, to guarantee high quality of the mutant samples, we chose to exclude them from further analysis. CNV data were obtained from the Affymetrix 6.0 platform and have 5 levels to represent different CNV status: deep deletion (CN = −2), copy loss (CN = −1), neutral (CN = 0), copy gain (CN = 1), and amplification (CN = 2). mRNA expression data from RNA-sequencing were downloaded in the form of RSEM values. For each gene, we normalized its expression against its copy neutral samples, following the strategy by cBio portal (Gao et al., 2013). Specifically, we transformed RSEM values into z-scores: $z = s - \mu/sd$, where s indicates the raw RSEM value for a gene in each sample, and μ and sd are the average and standard deviation of RSEM values in its copy neutral samples.

Statistical analysis of human TSGs

Test 1: Sample-based test to control per-sample mutation load: A TSG harboring L2 (or L1) mutations can be formulated using the Poisson binomial distribution. In each sample, the status of a gene forms the Bernoulli process: harboring a mutation of the investigated mutation type (the 'success' outcome) or not (the 'failure' outcome). Let $p_{i,L2} = \#\ genes\ with\ L2/n$ is the probability of L2 mutations in the i^{th} sample. Across samples, a gene with L2 (L1) mutations in K samples out of a total of N samples is the combination of multiple independent binomial distributions, each with its unique probability of success. We defined a score for each TSG in each cancer as $p_{TSG} = \prod_{i=1}^{K} p_i$. To assess the significance, we solve the problem in two steps. Step 1, to determine if K successes out of N trials is significantly different from random. The probability of L2 mutations in each cancer type, regardless of the sample variation, ranged between 6.202×10^{-4} (READ) and 0.106 (LAML), and the probability of L1 mutations ranged between 3.617×10^{-4} (PRAD) and 5.086×10^{-3} (COAD). In 58/62 cases, an observation of ≥ 5 mutations is not randomly expected using a binomial test ($p < 0.05$). Here, 62 is twice the number of cancer types (n = 33) with L2 and L1 tested separately for each cancer type, excluding four cases in which the tests were not conducted due to small sample size (L2 in KICH, L1 in MESO, L1 in THCA, and L2 in THYM). Thus, we used 5 as the minimum requirement for mutation events. Step 2, with the number of mutations (K) resolved, we next determined if the K successes were due to sample biases. We proposed a mutation load-matched resampling strategy to estimate the chance of observing the K samples with p, $i = 1, \ldots K$ for each TSG. Specifically, for each gene, we performed weighted resampling by selecting the same number of samples as K from N 10,000 times, where the chance for each sample to be selected is proportional to its L2 (L1) mutation load. In this way, a sample with a higher mutation load had a better chance to be selected, providing a simulation of the real cases where genes in such samples have a higher chance to harbor mutations. For each resampling, we calculate the probability of the gene, $p_{TSG}(\pi)$, in the same way as p_{TSG}. An empirical *p-value* was then calculated as $p_{sample} = \#\{p_{TSG}(\pi) > p_{TSG}\}/10000$. This test was conducted for inactivation mutations L2 and L1, respectively, in cancer types where the number of eligible samples was ≥ 20. Samples with

an exceptionally high mutation load were excluded from this sample-based test. We defined such hyper-mutated samples as those with mutations > upper quantile + 10 × IQR for L2 and L1, respectively. An exception was made for highly heterogeneous cancer types such as LUAD where more than 10 samples would have been excluded following the above criteria.

Test 2: Gene-based test to control gene-length related mutation variability: We fit a monotonically increasing smooth function to estimate the gene length effect: $y \sim gam(x)$, where y is the mutation frequency for a gene in a particular cancer type and x is the mRNA length. We also tried to use amino acid length as the predictor and the results were similar. After successful fitting, we calculated a chi-square value for the gene following the likelihood ratio $\chi^2 = 2|\ln(y_p/1 - y_p) - \ln(y/1 - y)|$, where y_p is the predicted frequency. The resulting *p-value* was denoted as p_{gene}.

Transcriptomic impact—For each TSG × cancer event, we randomized the labels for the inactivated samples and WT samples. In each randomized set, we conducted a differential expression analysis and calculated $T_{random} = \sqrt{t_{1,random}^2 + t_{2,random}^2}$. With 10,000 randomization trials, we calculated an empirical *p-value* (p_{emp}) for each TSG × cancer event as $p_{\text{emp}} = \#(T\text{random} > T)/10000$. p_{emp} measures the significance of the combined impacts of both *cis*- and *trans*-effects.

Filtering abnormal outliers at the transcriptomic level—In our manual inspection of mutant samples (i.e., those with L2 or L1 mutations), we observed occasional abnormal outliers in some genes, whose expression profile appeared as dissimilar to other mutation samples. For example, a TSG was reported with inactivation mutations but it showed no sign of decrease in its expression. In Figure S7A, we took *RB1* in the BRCA_Basal subtype as an example. There were 19 BRCA_Basal samples with *RB1* inactivation mutations, including 15 samples with deep deletion (L2), 3 samples with copy loss accompanied with truncation mutations (L2), and 1 sample with truncation mutations (L1). However, a visualization inspection revealed that two samples with deep deletion had quite high expression of *RB1*, implying these deep deletion events did not function as expected.

We then developed a strategy to systematically and quantitatively screen for such abnormal outliers. We used the combined impact score $T = \sqrt{t_1^2 + t_2^2}$ to determine whether a mutated sample is an outlier from the remaining of the mutated samples. Here t_1 is the *t*-value for the TSG itself (*cis*-effect) and t_2 is the average *t*-value for the top 1% DEGs in the transcriptome excluding the TSG itself (*trans*-effect). Hence, T measured an overall impact, instead of the TSG itself. Specifically, for each TSG in each cancer, we started with its L2+L1 samples and calculated a list of new T's, each corresponding to the exclusion of a mutated sample. If excluding a mutated sample would lead to an increase of T by 5%, i.e., $T' > T(1 + 0{:}05)$, then the corresponding mutated sample would be excluded from the L2+L1 sample pool. Notably, to avoid self-service analyses, such abnormal outlier samples were not re-grouped into the wild-type samples either but were permanently excluded from all following analyses. The process was iteratively repeated until no mutated sample was associated with an extreme T' that would increase the combined impact score by 5%.

With the quantitative assessment of potential outlier samples, the two samples with *RB1* in BRCA_Basal (Figure S7A) were identified as outlier samples, i.e., removing each of them could lead to an increase of the impact score T by > 5%. As shown in the right panel in Figure S7A, we observed a prominent increase in both the *cis*-effect and the *trans*-effect after removing these two outlier samples. Notably, the outlier samples were excluded from the inactivated samples permanently. They would not be categorized as WT samples and were not included in the following analysis.

We applied the filtering strategy for all 277 TSG × cancer events. As a result, 208/277 (75.1%) events remained unchanged, 53 (19.1%) were reduced by ≤ 3 samples, and 11 events were excluded due to insufficient samples (Figures S7B and S7C). For all the subsequent analysis, we used the 266 events with cleaned inactivated samples.

Pathway enrichment analysis—For each TSG × cancer event, we used the single sample gene set enrichment analysis (ssGSEA) method implemented in a R package *GSVA* (Hänzelmann et al., 2013). ssGSEA calculates an enrichment score (ES) for each pathway in each sample, resulting in an ES matrix with rows representing pathways and columns representing TCGA samples. The algorithms for calculation of ES can be found in Hänzelmann et al. (2013). ssGSEA was performed for each cancer type respectively. For each TSG × cancer event, we then used Wilcoxon Rank Sum test to compare the ES values in inactivated samples and the WT samples for the corresponding TSG for each pathway. The resultant *p-value*s were corrected for multiple testing using stringent Bonferroni method. Significant pathways (Bonferroni p value < 0.05) in > 16 TSG × cancer events (10% of all 161 events) were used to generate Figure 4E.

Impacted network—For each TSG in each cancer, we imposed the *t*-value of each gene onto a comprehensive human reference interaction network (denoted by G, with node size g). This network was built based on HPRD and String protein interaction databases. With nodes weighted by their *t*-values, we applied Random Walk with Restart (RWR) as follows: $p^{m+1} = r\boldsymbol{W}p^m + (1 - r)p^0$, where $r = 0.5$, p^0, p^m, and p^{m+1} are vectors of length g at time points 0, m, and $m+1$, and $\boldsymbol{W}$ is the column-normalized adjacency matrix for the network G. At time 0, each element in p^0 represents the *t*-value for the corresponding gene; and when the function converged (e.g., the difference between p^m and $p^{m+1} < 1 \times 10^{-3}$), the elements in the final p^{m+1} represent the probabilities that the walker would arrive at the corresponding nodes, referred to as $p^S = \left\{p_i^S\right\}$, $i = 1, \ldots g$. As a control, we initially applied RWR with all nodes equally weighted, i.e., $p_i^0 = 1/g$, $i = 1, \ldots g$. We referred to the results from this setting as the baseline probabilities, $p_{BASE} = \left\{p_i^B\right\}$, $i = 1, \ldots g$. For each TSG × cancer event, the probability of visiting each node in the network at the stable status was formulated as $z_i^B = p_i^S - mean(p_{BASE})/sd(p_{BASE})$. To further evaluate significance, we randomized the labels of inactivated samples and WT samples by 10,000 times, resulting in 10,000 sets of *t*-values. The resampled labels thus had no relationship with the TSG inactivation status. For each random set, we applied RWR following the above strategy, resulting in 10,000 random sets of p^S for each TSG, denoted by $p^{\pi} = \{p_{i,j}\}$, $i = 1, \ldots g$, $j = 1, \ldots, 10000$. For each gene, these probability values formed the null distribution and we calculated a second *z*-score as

$z_i^{\pi} = p_i^S - mean(p_i^{\pi})/sd(p_i^{\pi})$. We considered only those genes as impacted when they had $z_i^B \geq 2$ and $z_i^{\pi} \geq 2$, corresponding to 2 standard deviation from the center. Furthermore, only the genes whose shortest path to the corresponding TSG was ≤ 2 were included.

QUANTIFICATION AND STATISTICAL ANALYSES

Somatic mutation, CNV, and RNA-sequencing analyses were based on 7248 samples with qualified data. Definitions of significance for various statistical tests are described and referenced in their respective sections in the Method Details.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank Drs. Dung-Fung Lee, Jeffrey Chang, and Da Yang and Mr. Mi Li for insightful discussion. We also thank the two reviewers for their constructive comments and Dr. Irmgard Willcockson for English editing that improved the manuscript. We thank the TCGA Research Network for making the TCGA data available to the research community and the UTHealth Cancer Genomics Core and Data Science and Informatics Core for Cancer Research, supported by CPRIT (RP180734 and RP170668). This work was partially supported by an NIH grant (R01LM012806) and an American Cancer Society Institutional Research Grant (IRG-58-009-55). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, and Sunyaev SR (2010). A method and server for predicting damaging missense mutations. Nat. Methods 7, 248–249. [PubMed: 20354512]

Bowden GT, Schneider B, Domann R, and Kulesz-Martin M (1994). Oncogene activation and tumor suppressor gene inactivation during multistage mouse skin carcinogenesis. Cancer Res. 54, 1882s–1885s. [PubMed: 8137304]

Bowen C, Ju JH, Lee JH, Paull TT, and Gelmann EP (2013). Functional activation of ATM by the prostate cancer suppressor NKX3.1. Cell Rep. 4, 516–529. [PubMed: 23890999]

Chang X, and Wang K (2012). wANNOVAR: annotating genetic variants for personal genomes via the web. J. Med. Genet 49, 433–436. [PubMed: 22717648]

Chayka O, Corvetta D, Dews M, Caccamo AE, Piotrowska I, Santilli G, Gibson S, Sebire NJ, Himoudi N, Hogarty MD, et al. (2009). Clusterin, a haploinsufficient tumor suppressor gene in neuroblastomas. J. Natl. Cancer Inst 101, 663–677. [PubMed: 19401549]

Chen Z, Wang Z, Guo W, Zhang Z, Zhao F, Zhao Y, Jia D, Ding J, Wang H, Yao M, and He X (2015). TRIM35 Interacts with pyruvate kinase isoform M2 to suppress the Warburg effect and tumorigenicity in hepatocellular carcinoma. Oncogene 34, 3946–3956. [PubMed: 25263439]

Cheng F, Zhao J, Fooksa M, and Zhao Z (2016). A network-based drug repositioning infrastructure for precision cancer medicine through targeting significantly mutated genes in the human cancer genomes. J. Am. Med. Inform. Assoc 23, 681–691. [PubMed: 27026610]

Chun S, and Fay JC (2009). Identification of deleterious mutations within three human genomes. Genome Res. 19, 1553–1561. [PubMed: 19602639]

Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, and Sander C (2013). Emerging landscape of oncogenic signatures across human cancers. Nat. Genet 45, 1127–1133. [PubMed: 24071851]

Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A, and Sidow A; NISC Comparative Sequencing Program (2005). Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 15, 901–913. [PubMed: 15965027]

Davoli T, Xu AW, Mengwasser KE, Sack LM, Yoon JC, Park PJ, and Elledge SJ (2013). Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. Cell 155, 948–962. [PubMed: 24183448]

Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, and Batzoglou S (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput. Biol 6, e1001025. [PubMed: 21152010]

Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, and Liu X (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum. Mol. Genet 24, 2125–2137. [PubMed: 25552646]

Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci. Signal 6, pl1. [PubMed: 23550210]

Garber M, Guttman M, Clamp M, Zody MC, Friedman N, and Xie X (2009). Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics 25, i54–i62. [PubMed: 19478016]

Hänzelmann S, Castelo R, and Guinney J (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics 14, 7. [PubMed: 23323831]

Holbrook JA, Neu-Yilik G, Hentze MW, and Kulozik AE (2004). Nonsense-mediated decay approaches the clinic. Nat. Genet 36, 801–808. [PubMed: 15284851]

Hu Z, Yau C, and Ahmed AA (2017). A pan-cancer genome-wide analysis reveals tumour dependencies by induction of nonsense-mediated decay. Nat. Commun 8, 15943. [PubMed: 28649990]

Johnston JJ, Lewis KL, Ng D, Singh LN, Wynter J, Brewer C, Brooks BP, Brownell I, Candotti F, Gonsalves SG, et al. (2015). Individualized iterative phenotyping for genome-wide analysis of loss-of-function mutations. Am. J. Hum. Genet 96, 913–925. [PubMed: 26046366]

Knijnenburg TA, Wang L, Zimmermann MT, Chambwe N, Gao GF, Cherniack AD, Fan H, Shen H, Way GP, Greene CS, et al. (2018). Genomic and molecular landscape of DNA damage repair deficiency across The Cancer Genome Atlas. Cell Rep. 23, 239–254.e6. [PubMed: 29617664]

Knudson AG Jr. (1971). Mutation and cancer: statistical study of retinoblastoma. Proc. Natl. Acad. Sci. USA 68, 820–823. [PubMed: 5279523]

Kumar P, Henikoff S, and Ng PC (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc 4, 1073–1081. [PubMed: 19561590]

Man CH, Fung TK, Wan H, Cher CY, Fan A, Ng N, Ho C, Wan TS, Tanaka T, So CW, et al. (2015). Suppression of SOX7 by DNA methylation and its tumor suppressor function in acute myeloid leukemia. Blood 125, 3928–3936. [PubMed: 25940713]

Mavrakis KJ, McDonald ER 3rd, Schlabach MR, Billy E, Hoffman GR, deWeck A, Ruddy DA, Venkatesan K, Yu J, McAllister G, et al. (2016). Disordered methionine metabolism in MTAP/CDKN2A-deleted cancers leads to dependence on PRMT5. Science 351, 1208–1213. [PubMed: 26912361]

Mizuarai S, Machida T, Kobayashi T, Komatani H, Itadani H, and Kotani H (2011). Expression ratio of CCND1 to CDKN2A mRNA predicts RB1 status of cultured cancer cell lines and clinical tumor samples. Mol. Cancer 10, 31. [PubMed: 21447152]

Møller MB, Ino Y, Gerdes AM, Skjødt K, Louis DN, and Pedersen NT (1999). Aberrations of the p53 pathway components p53, MDM2 and CDKN2A appear independent in diffuse large B cell lymphoma. Leukemia 13, 453–459. [PubMed: 10086736]

Reva B, Antipin Y, and Sander C (2011). Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 39, e118. [PubMed: 21727090]

Ryland GL, Doyle MA, Goode D, Boyle SE, Choong DY, Rowley SM, Li J, Bowtell DD, Tothill RW, Campbell IG, and Gorringe KL; Australian Ovarian Cancer Study Group (2015). Loss of heterozygosity: what is it good for? BMC Med. Genomics 8, 45. [PubMed: 26231170]

Schwarz JM, Rödelsperger C, Schuelke M, and Seelow D (2010). MutationTaster evaluates disease-causing potential of sequence alterations. Nat. Methods 7, 575–576. [PubMed: 20676075]

Schweingruber C, Rufener SC, Zünd D, Yamashita A, and Mühlemann O (2013). Nonsense-mediated mRNA decay - mechanisms of substrate mRNA recognition and degradation in mammalian cells. Biochim. Biophys. Acta 1829, 612–623. [PubMed: 23435113]

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, and Ideker T (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13, 2498–2504. [PubMed: 14597658]

Shapiro GI, Edwards CD, Kobzik L, Godleski J, Richards W, Sugar-baker DJ, and Rollins BJ (1995). Reciprocal Rb inactivation and p16INK4 expression in primary lung cancers and cell lines. Cancer Res. 55, 505–509. [PubMed: 7834618]

Shihab HA, Gough J, Cooper DN, Day IN, and Gaunt TR (2013). Predicting the functional consequences of cancer-associated amino acid substitutions. Bioinformatics 29, 1504–1510. [PubMed: 23620363]

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, and Mesirov JP (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA 102, 15545–15550. [PubMed: 16199517]

Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr., and Kinzler KW (2013). Cancer genome landscapes. Science 339, 1546–1558. [PubMed: 23539594]

Zhao M, Sun J, and Zhao Z (2013). TSGene: a web resource for tumor suppressor genes. Nucleic Acids Res. 41, D970–D976. [PubMed: 23066107]

Zhao M, Kim P, Mitra R, Zhao J, and Zhao Z (2016). TSGene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. Nucleic Acids Res. 44, D1023–D1031. [PubMed: 26590405]

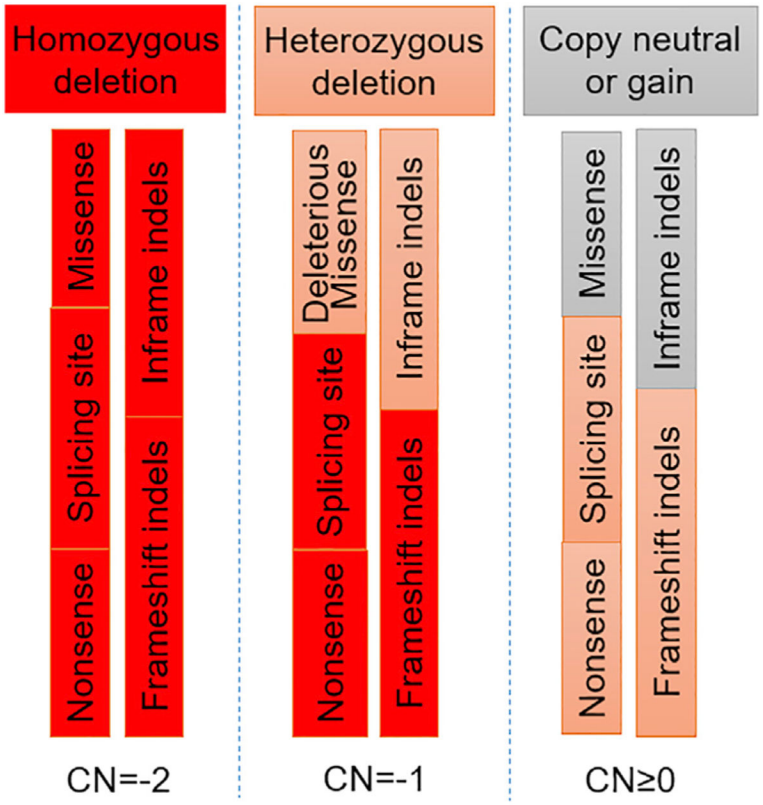

Figure 1. Definition of Inactivation Events

L2 mutations (in red): both copies were lost because of genetic mutations. L1 mutations (in orange): strong evidence for loss of only one copy, whereas the other copy has less-compelling evidence of disruption. Mutation types in gray do not have evidence of the loss of two copies.

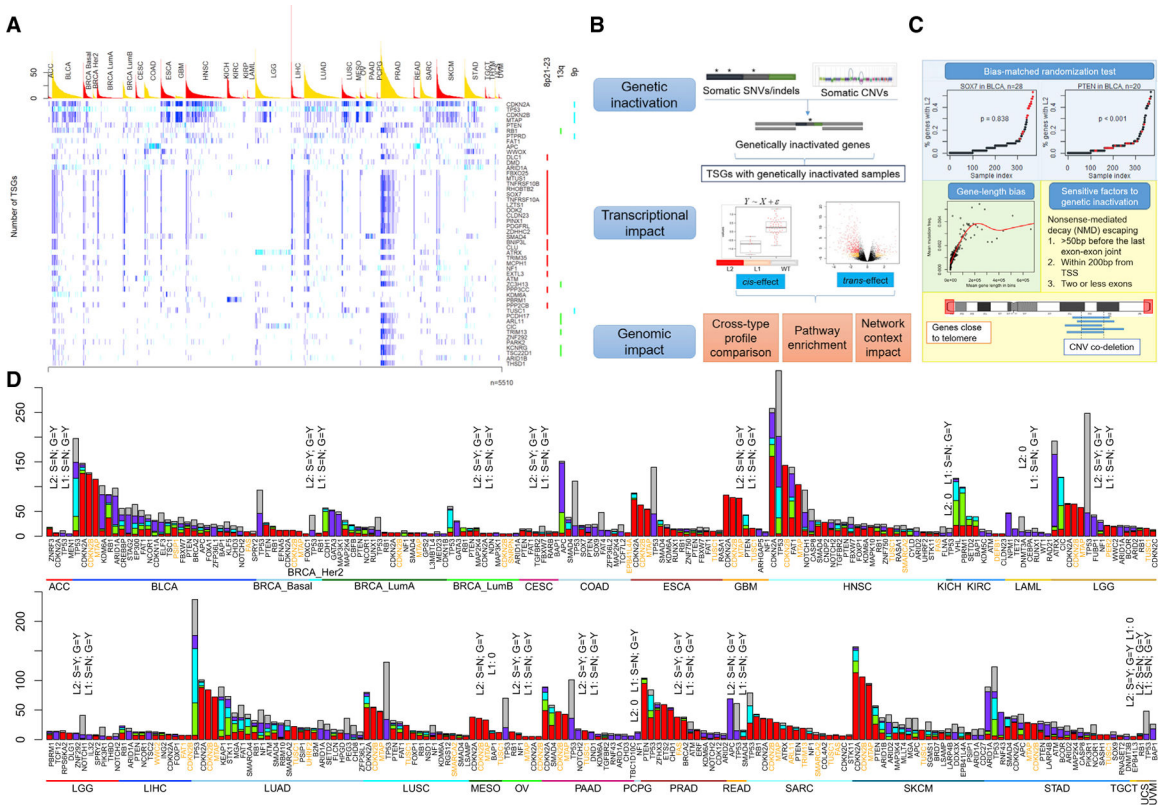

Figure 2. Identification of TSG Inactivation Events in 33 Cancer Types/Subtypes

(A) Overview of genetic inactivation events of tumor-suppressor genes (TSGs) (without any statistical test). This shows genetic inactivation events of the top-40 TSGs that had the highest frequency of inactivation events in 32 cancer types or subtypes (THCA was excluded, because no TSGs had ≥ 5 mutations). Vertical bars in blue and cyan label the samples with L2 and L1 mutations of a TSG (rows), respectively.

(B) Computational framework for identifying TSG inactivation events and their follow-up biological and functional impact analysis. First, we defined the genetically inactivated samples and WT samples for each TSG. Second, we examined the impact of the inactivation events on the gene expression of the TSG itself (*cis*-effect) and other genes in the transcriptome (*trans*-effect). Third, we studied other genomic impacts of the TSG inactivation events, including cross-type transcriptional profile similarity, pathways, and local networks.

(C) Summary of the computational pipeline to detect genetic inactivation events in TSGs.

(D) Overview of all TSG × cancer events. y axis indicates sample size. The colors label different types of mutations. Definition of mutation types: Homdel (red), homozygous deletion by CNVs (CNV = −2); Hetlose_NS_FS (green), heterozygous deletion by CNVs (CNV = −1) plus nonsense SNVs or frameshift indels; Hetlose_Mis (cyan), heterozygous deletion by CNVs (CNV = −1) plus missense SNVs; WT_NS_FS (purple): nonsense SNVs or frameshift indels occurred in samples without copy loss (CNV ≥ 0); WT_Mis (gray), missense SNVs occurred in samples without copy loss (CNV ≥ 0). Genes whose names are in orange were co-located with other genes in the human genome. For either L2 or L1, we applied a sample-based test and gene-length test wherever applicable (i.e., eligible sample

sizes ≥ 20; # mutated genes ≥ 100). When such criteria could not be met, the sample test was skipped (denoted by L2: S = N or L1: S = N) or the gene test was skipped (denoted by L2: G = N or L1: G = N).

ACC, adrenocortical carcinoma; BLCA, bladder urothelial carcinoma; BRCA, breast invasive carcinoma; CESC, cervical squamous cell carcinoma and endocervical adenocarcinoma; CHOL, cholangiocarcinoma; CESC, cervical squamous cell carcinoma; COAD, colon adenocarcinoma; ESCA, esophageal carcinoma; GBM, glioblastoma multiforme; HNSC, head and neck squamous cell carcinoma; KICH, kidney chromophobe; KIRC, kidney renal clear cell carcinoma; KIRP, kidney renal papillary cell carcinoma; LAML, acute myeloid leukemia; LGG, brain lower grade glioma; LIHC, liver hepatocellular carcinoma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; MESO, mesothelioma; OV, ovarian serous cystadenocarcinoma; PAAD, pancreatic adenocarcinoma; PCPG, pheochromocytoma and paraganglioma; PRAD, prostate adenocarcinoma; READ, rectum adenocarcinoma; SARC, sarcoma; SKCM, skin cutaneous melanoma; STAD, stomach adenocarcinoma; TGCT, testicular germ cell tumors; THCA, thyroid carcinoma; THYM, thymoma; UCEC, uterine corpus endometrioid carcinoma; UCS, uterine carcinosarcoma; UVM, uveal melanoma.

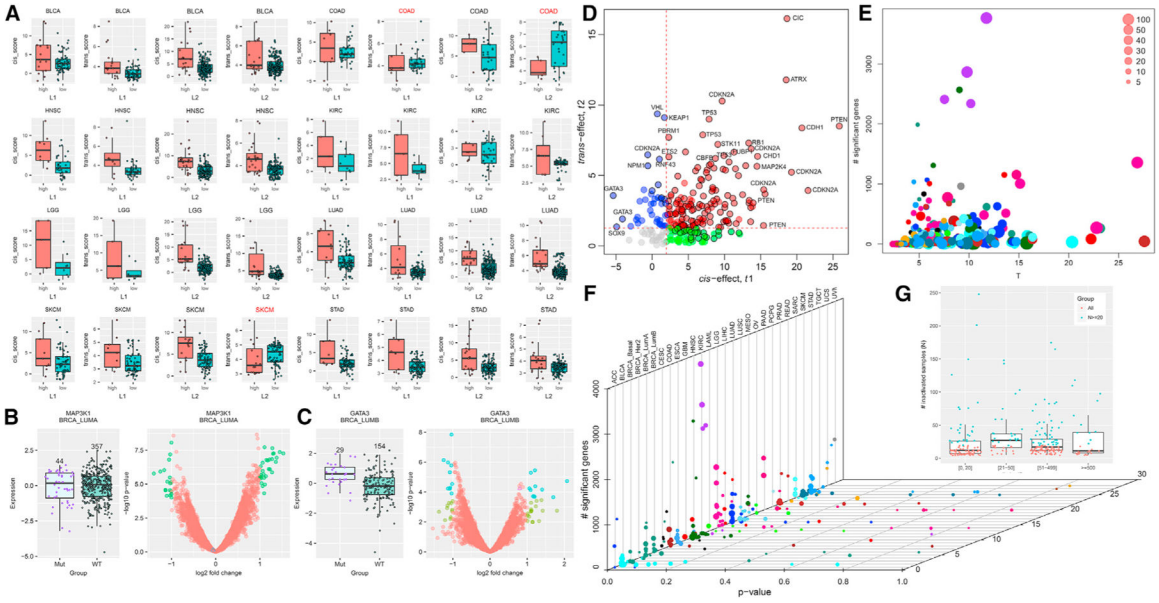

Figure 3. Inactivation Events of TSGs with Cis- and Trans-Effects

(A) Evaluation of TSG × cancer inactivation events using transcriptomic data. In each panel, the group labeled "high" indicates the significant TSG × cancer events ($p_{sample} < 0.05$ and $p_{gene} < 0.05$), and the group "low" indicates non-significant events. *Cis*- and *trans*-effects were assessed for L1 mutations and L2 mutations, respectively.

(B and C) *MAP3K1* (B) and *GATA3* (C) as example genes that had no *cis*-effect but had a *trans*-effect. In both cases, as shown, although *MAP3K1* and *GATA3* were both found with inactivation mutations, their mRNA expression level was not decreased compared to the WT samples of the corresponding genes.

(D) Plot of all TSGs for their *cis*- and *trans*-effects. Each dot indicates a TSG × cancer event. Node color indicates the events with different *cis*- or *trans*-effects. Nodes with a black circle indicate the 161 events with significant impact.

(E) Distribution of DEGs (y axis) versus T (see main text for the definition of T). Each dot represents a TSG × cancer event. Node color indicates cancer types (like in F), and node size is proportional to the number of inactivated samples.

(F) Distribution of differentially expressed genes (DEGs, y axis) versus empirical p_{emp} (x axis) of each TSG in each cancer type. Color for cancer types are the same as those in (E).

(G) Distribution of significant DEGs.

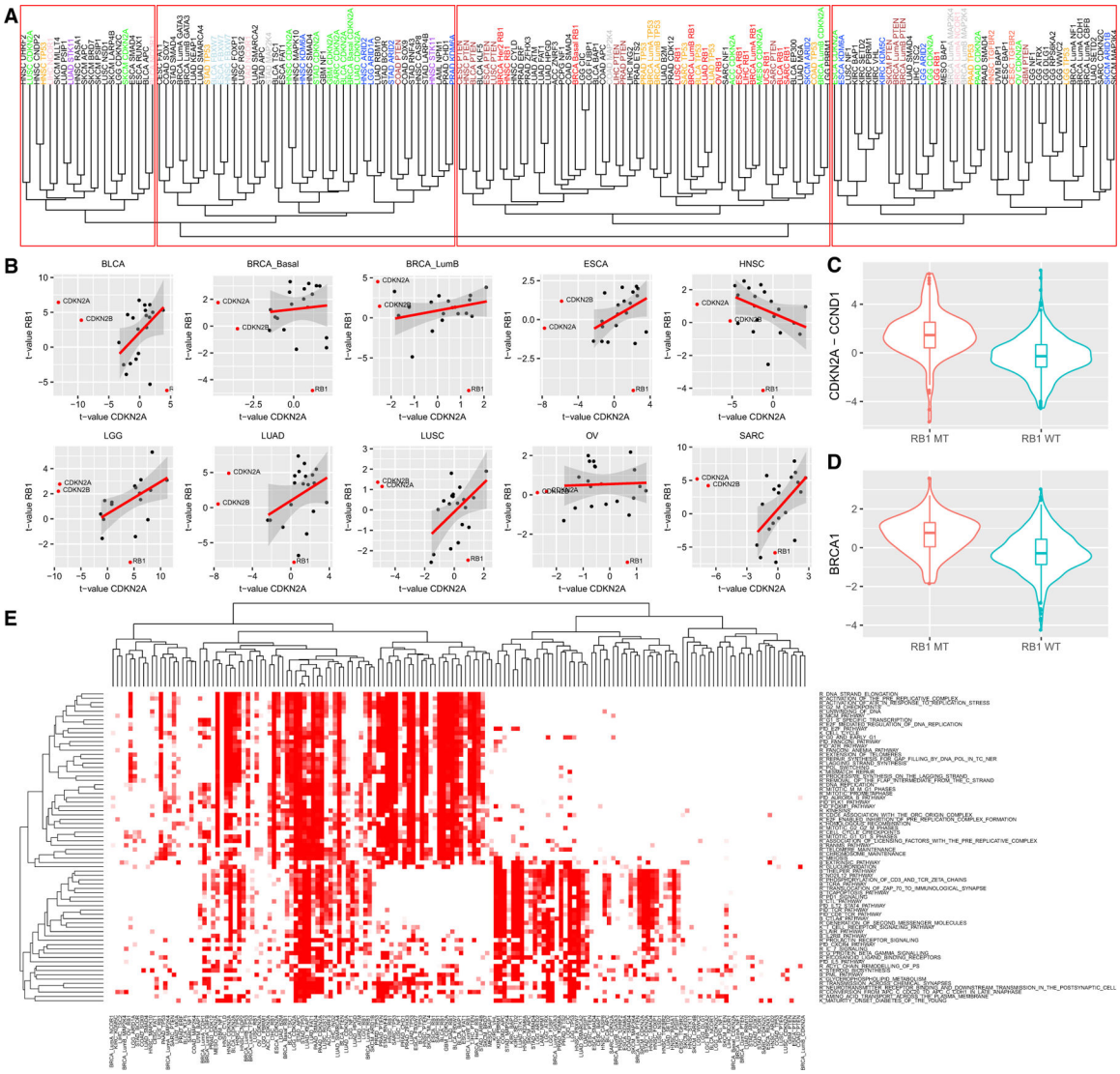

Figure 4. Transcriptome Impact of TSG Inactivation Events

(A) Hierarchical cluster of 161 TSG × cancer events. The dendrogram was obtained based on a 10,000 bootstrap (see main text).

(B) Comparison of cell-cycle genes in *RB1*-inactivated samples and in *CDKN2A*-inactivated samples.

(C) Gene expression of *CDKN2A* and *CCND1* in *RB1*-inactivated samples and wild-type samples and restricted to *CDKN2A* wild-type samples.

(D) Gene expression of *BRCA1* in *RB1*-inactivated samples and wild-type samples and restricted to *CDKN2A* wild-type samples.

(E) ssGSEA results using mRNA expression.

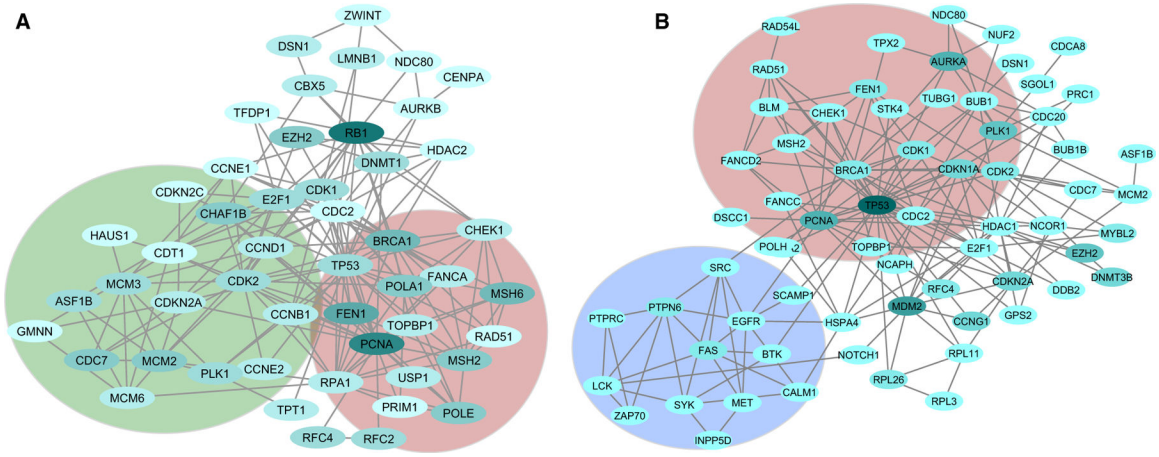

Figure 5. Affected Local Networks by RB1 and TP53 Inactivation Events

(A) Affected network for *RB1*. Recurrent nodes that were significantly changed in ≥ 5 cancer types were selected to plot on the graph. Node color is proportional to the number of cancer types in which the gene was changed. The nodes shaded in green mainly function in the cell-cycle pathway, and the nodes shaded in pink are mainly from the DNA repair pathway.

(B) Affected network for *TP53*. Recurrent nodes that were significantly changed in ≥ 2 cancer types were selected to plot on the graph. The shaded areas were for signaling pathway (blue) and cell-cycle pathway (red), respectively.

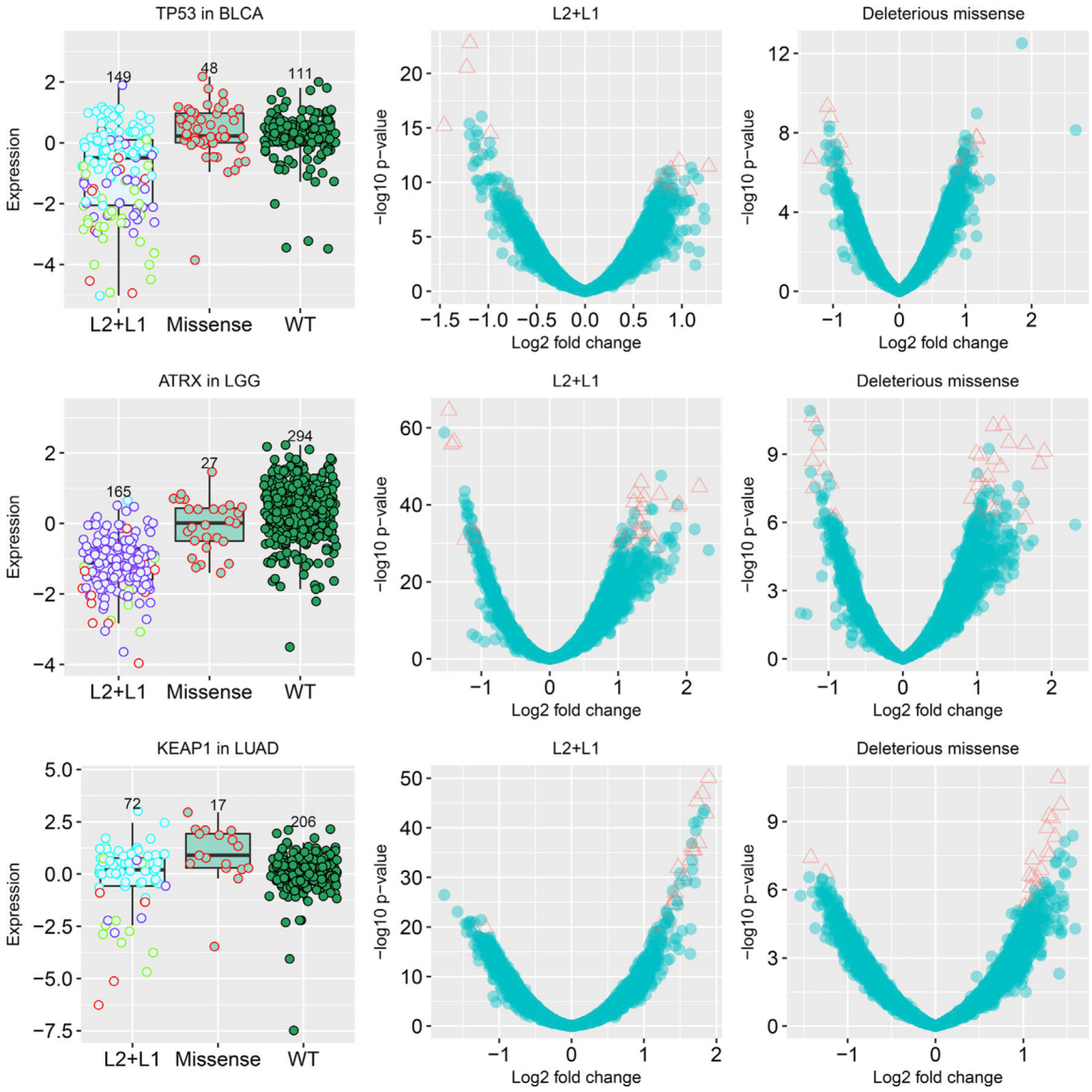

Figure 6. Exploring the Impact of Missense Mutations on TSG Function

Three example genes (*TP53*, *ATRX*, and *KEAP1*), whose deleterious missense mutations had an effect similar to other inactivation mutations, are shown. Triangle dots indicate genes that were detected as DEGs in both comparisons: they were DEGs when comparing the inactivated samples with the WT samples and were also DEGs when comparing the samples harboring missense mutations (gray bars in Figure 2D) with the WT samples. Note that those samples with missense mutations had no overlap with the inactivated samples.

KEY RESOURCES TABLE

Reagent or resource	Source	Identifier
Deposited Data		
TCGA somatic mutation data	UCSC Cancer Genome Browser Xena	https://xena.ucsc.edu/
TCGA somatic copy number data	UCSC Cancer Genome Browser Xena	https://xena.ucsc.edu/
TCGA gene expression data	UCSC Cancer Genome Browser Xena	https://xena.ucsc.edu/
Tumor Suppressor Genes	Zhao et al., 2013; Zhao et al., 2016	https://bioinfo.uth.edu/TSGene/
Tumor Suppressor Genes	Davoli et al., 2013	TUSON
Software and algorithms		
Single sample gene set enrichment analysis (ssGSEA)	Hänzelmann et al., 2013	Hänzelmann et al., 2013
Cytoscape	Shannon et al., 2003	https://cytoscape.org/