Molecular Therapy
Methods & Clinical Development
Original Article

Identification and Characterization of the Copy Number Dosage-Sensitive Genes in Colorectal Cancer

Zhiqiang Chang,[1,2] Xinxin Liu,[1,2] Wenyuan Zhao,[1] and Yan Xu[1]

[1]College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China

Dosage effect is one of the common mechanisms of somatic copy number alteration in the development of colorectal cancer, yet the roles of dosage-sensitive genes (DSGs) in colorectal cancer (CRC) remain to be characterized more deeply. In this study, we developed a five-step pipeline to identify DSGs and analyzed their characterization in CRC. Results showed that our pipeline performed better than existing methods, and the result was significantly overlapped between solid tumor and cell line. We also found that the top five DSGs (*PSMF1, RAF1, PTPRA, MKRN2*, and *ELP3*) were associated with the progression of CRC. By analyzing the characterization, DSGs were enriched in driver genes and they drove sub-pathways of CRC. In addition, immune-related DSGs are associated with CRC progression. Our results also showed that the CRC samples affected by high microsatellites have fewer DSGs, but a higher overlap with DSGs in microsatellite low instability and microsatellite stable samples. In addition, we applied DSGs to identify potential drug targets, with the results showing that 22 amplified DSGs were more sensitive to four drugs. In conclusion, DSGs play an important role in CRC, and our pipeline is effective to identify them.

INTRODUCTION

Colorectal cancer (CRC) is the third leading cause of cancer-associated deaths in the world.[1] Studies have shown that somatic copy number alteration (SCNA) is one of the most common structural variations in CRC.[2,3] SCNA is defined as the phenomenon of amplifying or deleting genome fragments, which widely exist in the human cancer genome, and it is also usually considered to be the driver event and development of CRC.[4] Existing studies have shown that SCNA affects development and other functions,[5] and different levels of SCNA may have adverse effects on cancers.[6,7] Thus, SCNA could also be used as an important prognostic marker for cancer.[8,9] Taken together, the results of these studies illustrated that SCNA plays an important role in CRC.

Dosage effect is one of the most important mechanisms[10–12] by which SCNA disrupts the gene function and causes an abnormal phenotype. If the expression of a gene in the region of the SCNA increases with the SCNA value, and vice versa, it could be defined as a dosage-sensitive gene (DSG).

It is well known that gene expression is regulated by transcription factors, microRNAs (miRNAs), long non-coding RNAs (lncRNAs), and other regulatory elements,[13] which are not always upregulated with the increase in copy number. The dosage sensitivity of the copy number is an important complementary mechanism for gene transcription regulation. Therefore, the identification of DSGs is highly important in CRC studies.

Recent studies have discovered DSGs in CRC. Komor et al.[14] reported that *POFUT1*, *RPRD1B*, and *EIF6* were observed by DNA copy number-driven gene-dosage effect in high-risk CRC, and *POFUT1* was considered as a candidate driver of CRC progression. The computing method has been developed to identify DSGs. According to the definition of DSGs, linear regression methods have been widely used.[15,16] Samur et al.[17] developed a method that computed a qualitative dosage effect scoring. In addition, Yan et al.[18] developed a pipeline by combining the rank relationship of the copy number among genes and the linear relationship between SCNA and gene expression. The core idea of these methods was still a linear relationship. In an opinion, Veitia et al.[19] pointed out that the relationship between the gene expression and copy number showed a non-linear behavior. However, we still knew little about which non-linear function or curve could describe this non-linear relationship. Therefore, an effective pipeline to represent this relationship needed to be proposed and proven to be effective.

Based on the importance of SCNA in CRC and the insufficient identifying method of DSGs, we developed an effective method called PDSG to identify DSGs in CRC.

RESULTS

Comparison with Existing Methods

Linear regression is one of the most commonly used pipelines for identifying DSGs.[15,16] In order to test whether the pipeline for identifying a DSG (PDSG) was superior to the linear regression pipeline, we separately applied the two pipelines to calculate the dosage effect score in CRC. The results showed that PDSG had a higher dosage effect score (Figure 1A) and a lower sum of squares of residuals

Received 29 February 2020; accepted 22 June 2020;
https://doi.org/10.1016/j.omtm.2020.06.020.

[2]These authors contributed equally to this work.

Correspondence: Yan Xu, College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China.
E-mail: xuyan@ems.hrbmu.edu.cn

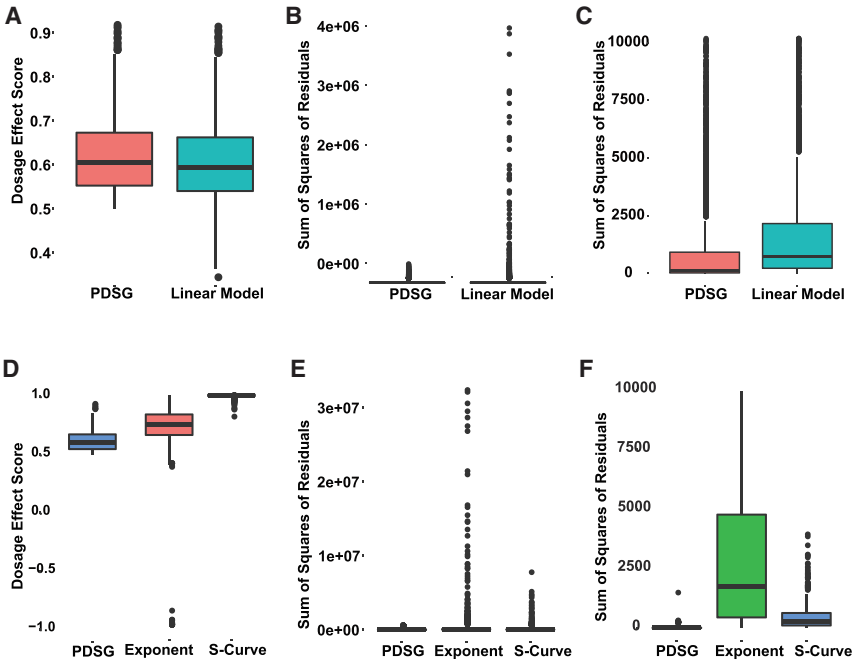

Figure 1. Results of the Different DSG-Identifying Methods

(A) The dosage effect score of PDSG and the linear model. (B and C) The sum of squares of residuals of PDSG and the linear model. (B) The overview of the sum of squares of residuals, and (C) was a part of (B). (D) The dosage effect score of PDSG, exponent, and S-curve. (E and F) The sum of squares of residuals of PDSG, exponent model and S-curve model. (E) Overview of the sum of squares of residuals, and (F) was a part of (E).

(Figures 1B and 1C). These findings indicated that PDSG was more suitable than linear regression to describe the relationship between SCNA and gene expression.

In order to test whether the performance of PDSG was better than other non-linear models in identifying DSGs, we compared it with the two modes of exponential transformation and S-curve transformation. The dosage effect scores of the two comparison models were higher than PDSG (Figure 1D), but they had an abnormally greater sum of squares of residuals (Figures 1E and 1F). By analyzing fitting data, we found that the exponential function model transformed the expression data into super-large data sets, which caused the results to have a super-great sum of squares of residuals. Fitting of the S-curve produces most data that closed to the 0 value and super-elevation value, which may be due to the high value and low value parts of the S-curve at the same time. This led to the results of having a high sum of squares of residual as well. Compared with these nonlinear models, PDSG was thus more reasonable and appropriate for the relationship between SCNA and gene expression.

DSGs in CRC

We applied the PDSG to calculate the dosage effect score of each gene in CRC from The Cancer Genome Atlas (TCGA). A total of 10,860 genes were identified, and the statistics of the dosage effect scores of the genes are listed in Table S1. Then, we used PDSG to identify DSGs from the cell line dataset. Using the thresholds 0.5, 0.6, 0.7, and 0.8, we found that both results from solid tumor and the cell line had a significant overlap with each other (Figure 2A; hyper-geometric test, $p < 0.01$). Figure 2B showed the scores of the top five DSGs (*PSMF1*, *RAF1*, *PTPRA*, *MKRN2*, and *ELP3*) in solid tumor and cell line data in which the average dosage effect score was 0.76, significantly higher than that at random (Figure 2C; disturbance experiment, $p < 0.05$). These indicated that PDSG was robust in identifying DSGs in CRC.

The results above showed that our PDSG was better than other pipelines or methods in identifying DSGs, and it was robust in different datasets. Because the number of CRC patients with SCNA in TCGA is large enough and the number of those in the Cancer Cell Line Encyclopedia (CCLE) is not enough to analyze, we therefore analyzed the DSGs from TCGA in following study instead of analyzing the common genes between the results from TCGA and CCLE.

In order to investigate whether DSGs with a higher dosage effect score affected the progress of CRC, we performed a literature review of the top five DSGs. The results showed that all of them were directly or indirectly related to CRC. The gene *PSMF1* is an antagonistic regulator of the proteasomal activity,[20] while silencing *Rpt4* reduced the proteasomal activity in CRC.[21] The gene *RAF1* played a key role in maintaining the transformation phenotype of CRC cells.[22] Dephosphorylated *RAF1* reduced the signal transduction, increased the invasion and migration activity of CRC cells, and activated the epithelial-mesenchymal transformation.[23] *MKRN2* was found to be significantly upregulated in the CRC cell lines.[24] Finally, *ELP3* was downregulated in the colorectal oxaliplatin-resistant cell lines.[25] These facts supported the suggestion that DSGs with higher dosage effect scores played more important roles in CRC.

To further examine whether the DSGs with the highest dosage effect score were associated with prognosis in CRC patients, we performed a survival analysis (Figure 2D), which showed that the deletion of the gene *PSMF1* was significantly correlated with the disease-specific survival time in patients (log-rank test, $p = 3.8e{-}2$), and the deletion of the genes *RAF1* and *MKRN2* was significantly correlated with the progression-free interval in the patients (log-rank test, $p = 3.85e{-}3$ and $p = 1.96e{-}2$, respectively). Using the independent data of CRC (GEO: GSE75500), the Kaplan-Meier curves of the genes *RAF* and *MKRN2* showed that the survival times in the samples of copy number deletion also had the same trend with those in TCGA (Figure S1).

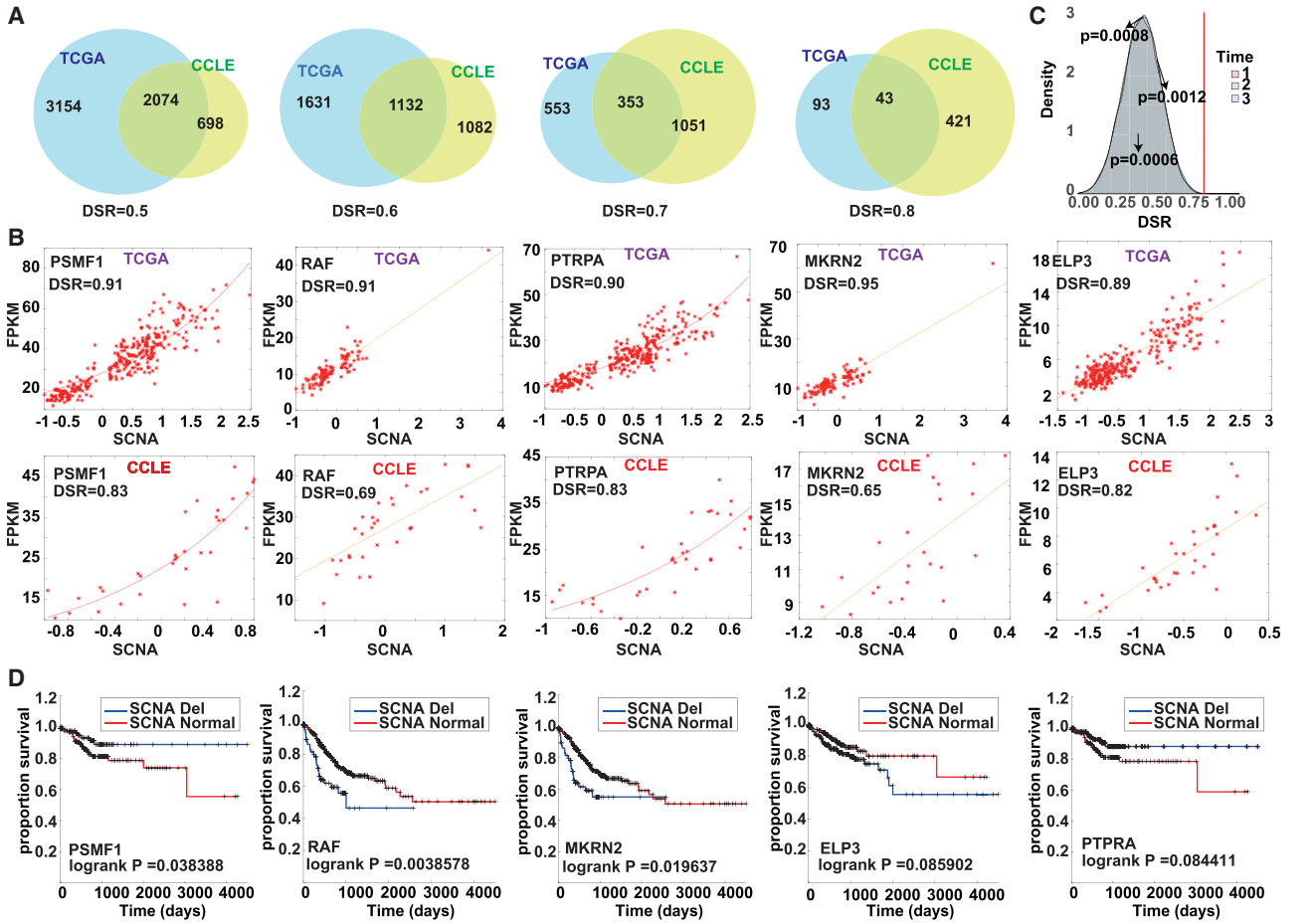

Figure 2. Overview of the DSGs in CRC

(A) Overlap of DSGs between solid tumor data from TCGA and cell line data from CCLE with the dosage-sensitive relationship thresholds of 0.5, 0.6, 0.7, and 0.8. (B) Dosage-sensitive relationship scores in TCGA and CCLE. (C) Results of the disturbance experiment, indicating that the DSR of the top five genes is significantly higher than that in random genes. (D) Kaplan-Meier curves of the five DGSs with top DSR.

Although the significant p values of the genes *PTRPA* and *ELP3* were not lower than 0.05, the deletion of *PTPRA* and *ELP3* was significantly correlated with the disease-specific survival time of the patients (log-rank test, p = 8.44e−2 and p = 8.59e−2, respectively), indicating that the copy number deletion of these two genes had a certain indicative effect on the prognosis of the patients. These results indicated that the DSGs with a higher dosage effect score had a closer relationship with the prognosis of the patients.

Effect of Microsatellite Stability on DSGs

In order to further investigate the influence of microsatellite stability (MSS) on DSGs, according to the provided microsatellite information, CRC samples were divided into the following three categories: MSS, microsatellite high instability (MSI-H), and microsatellite low instability (MSI-L). The results (Figure 3A) showed that the proportion of SCNA mutations in MSI-H samples is significantly lower than that of MSS and MSI-L samples.

When comparing their dosage sensitivity scores, statistical analysis (Figure 3B) showed no significant difference in dosage sensitivity scores between MSI-L and MSS samples (Wilcoxon rank-sum test, p = 136.90e−003). The sensitivity score of the genes in MSI-H samples was significantly lower than that of MSI-L and MSS samples (Wilcoxon rank-sum test, p = 60.59e−102 and 244.15e−111). When the filtered dosage sensitivity score is greater than 0.5, the results (Figure 3C) show that although the number of DSGs in MSI-H is less than in MSS, the DSGs of MSI-H have a high overlap with MSI-L and MSS (90.86% and 93.71%, respectively), indicating that most of the DSGs in MSI-H samples also showed higher dosage sensitivity in MSI-L and MSS samples. Functional analysis showed that the overlap genes are mainly enriched in important biological functions of cells, including non-coding RNA (ncRNA) processing, ribosome biogenesis, rRNA metabolic process, ubiquitin ligase complex, endosomal sorting complexes required for transport (ESCRT) complex, and nuclear envelope organization (Figure 3D).

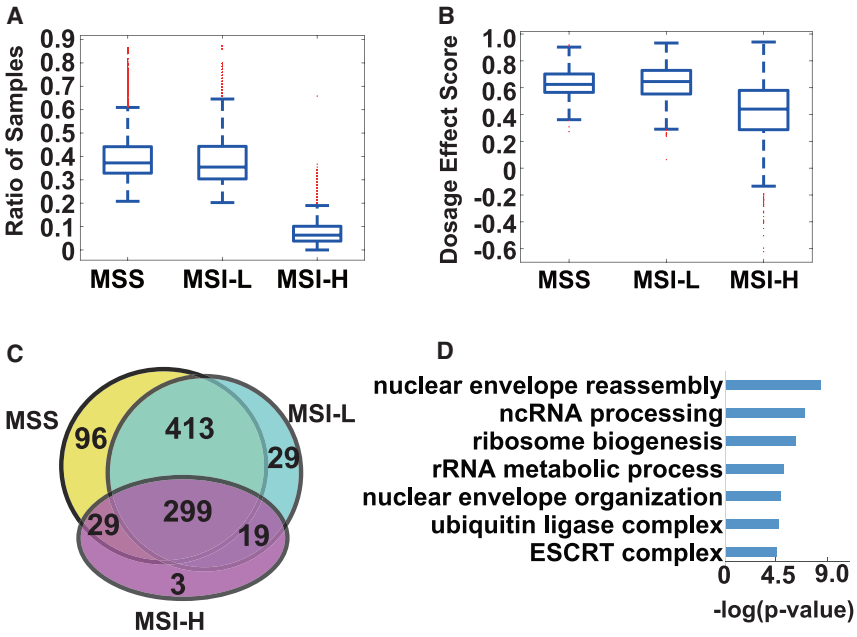

Figure 3. Microsatellite Analysis of CRC

(A) Proportion of SCNA samples from different microsatellite states, of which MSI-H samples have a relatively low alteration ratio. (B) Dose-dosage sensitivity score in different microsatellite states, of which MSI-H samples have a lower score. (C) Venn diagram of DSGs under different microsatellite states. MSI-H has the least DSGs, but it has a high overlap with MSS and MSI-L. (D) Function enrichment of the overlap DSGs among MSS, MSI-H, and MSI-L samples.

DSGs Drive the Pathway of CRC

By mapping the DSGs with a higher dosage effect score (>0.5) to the list of driver genes,[26] 30 genes were found, and they significantly overlapped with known driver genes (hyper-geometric test, p = 1.11e−2). In order to investigate whether they affected or drove the pathway of CRC, the DSGs with higher dosage effect score from TCGA and CCLE were mapped to the CRC pathway in the Kyoto Encyclopedia of Genes and Genomes (KEGG). The results showed that 20 and 11 DSGs in TCGA and CCLE were pathway-specific, respectively, and the overlap of these two datasets consisted of 10 genes, also indicating a higher consistency between the solid tumor and cell line.

In the CRC pathway, nearly every sub-pathway had at least more than one DSG with a higher dosage effect score. The genes *MAPK1*, *MAP2K1*, *MAPK8*, *CASP9*, *CASP3*, *BAD*, and *BAX* also existed in more than one sub-pathway. Furthermore, the genes in sub-pathway 8 (*EGFR*, *GRB2*, *SOS1*, *HRAS*, *MAP2K1*, *MAPK1*, *MTOR*, and *RPS6KB1*) and the genes in sub-pathway 4 (*RALGDS*, *RALA*, *RHOA*, and *MAPK8*) (Figure 4) were all DSGs, indicating that DSGs played an important role in the CRC pathway. In addition, the star mark in the CRC pathway showed that six DSGs in sub-pathways 8 and 4 have been marked with "survival" in the KEGG CRC pathway. To further examine whether the SCNA of these genes was associated with patient survival, log-rank test analysis showed that the deletion of *MTOR*, *MAPK8*, and *ROHA* and the amplification of *RALGDS* are associated with a poor prognosis (Figure 4), while the amplification of *RPS6KB1* was associated with a better prognosis (log-rank test, p = 1.6e−2). The results provided a complement to the CRC pathway, and we inferred that the dosage sensitivity of these five genes was a possible mechanism that influences the prognosis of patients, providing a new theoretical basis for our understanding of the CRC pathway.

Immuno-related DSGs Are Associated with Prognosis

With the purpose of testing whether DSGs could affect the immune response, we retrieved 39 immunosuppressive genes from the work of Zhang et al.[27] By mapping them to the DSGs with a higher dosage effect score, five DSGs were discovered, including *CD47*, *SPATA2*, *STAT3*, *VEGFA*, and *LGALS3*.

Studies have shown that the deletion of *CD47* by promoting angiogenesis promotes tumor progression,[28] and it inhibited the activation of cell cycle inhibitors and activated he expression of its promoters, which sped up the cell cycle process.[29] As a result, we inferred that the patients with the deletion of *CD47* copy number would have a poor prognosis. Consistent with what we inferred, the log-rank test showed that the samples with *CD47* copy number deletion had a lower survival time than did wild-type samples (Figure 5; log-rank test, p = 2.6e−2), and the same trend was found in the independent dataset of GEO: GSE75500 in CRC (Figure S2).

Vascular endothelial growth factor A (*VEGFA*) is a member of the *PDGF/VEGF* growth factor family. This growth factor plays a role in angiogenesis and endothelial cell growth, and it can induce endothelial cell proliferation, promote cell migration, inhibit apoptosis, and induce vascular permeability, which are necessary for both physiological and pathological angiogenesis. Silencing or downregulation of *VEGF* promotes cell apoptosis.[30,31] Thus, we inferred that the overexpression of *VEGFA* inhibits apoptosis, that the amplification of *VEGFA* could lead to the upregulation of its expression, and that the patients with amplified *VEGFA* would have a poor prognosis. Survival analysis revealed that the samples with *VEGFA* amplification had a poor prognosis (Figure 5; log-rank test, p = 4.19e−2).

STAT3 is known to play a carcinogenic role in a variety of malignant tumors, and the study of Grabner et al.[32] has shown that *STAT3* knockout could increase tumor growth, possibly due to the destruction of the immune function of *STAT3* itself. Clinically, low expression of *STAT3* was also associated with a poor prognosis. Based on this fact, we inferred that the amplification of *STAT3* might increase its expression level and have an anti-tumor effect to some extent. Prognostic analysis showed that the *STAT3* copy number amplified

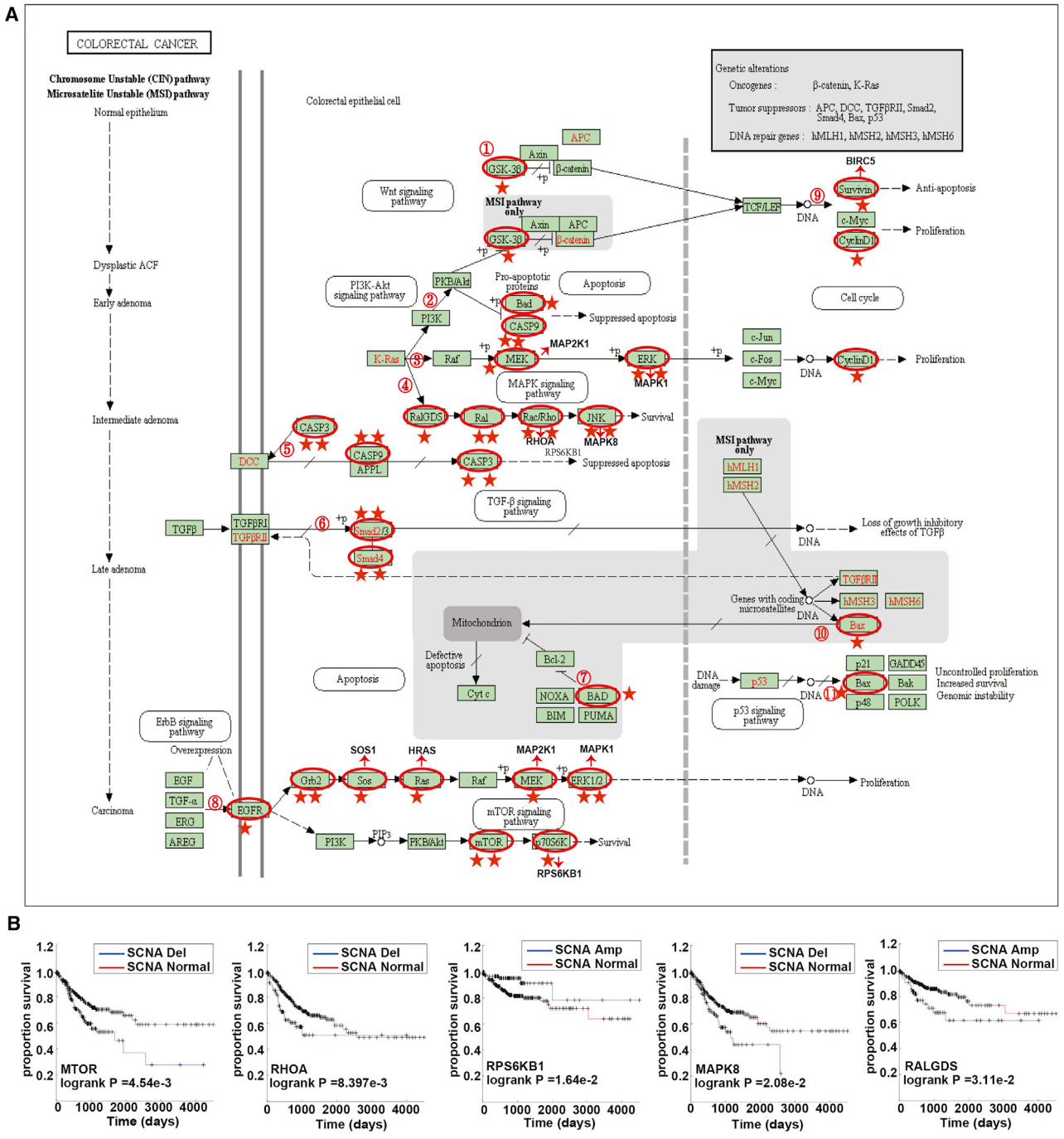

Figure 4. DSGs That Are Driver Genes of the CRC Pathway

(A) Mapping of DSGs in the CRC pathway from KEGG. The genes with a star mark represent DSGs, with one star mark representing the genes that were dosage-sensitive in either the data in solid tumor or the data in the cell line, and two star marks representing the genes in both the solid tumor and the cell line. (B) Kaplan-Meier curves of five DGSs in the CRC pathway from KEGG.

samples had a relatively low risk of prognosis (disease-specific survival) compared with the normal copy number samples (Figure 5; log-rank test, p = 3.66e−2), the same trend was found in the independent dataset of GEO: GSE75500 in CRC (see Figure S2). These results suggested that dosage-sensitive immunosuppressive genes affected the progression of CRC.

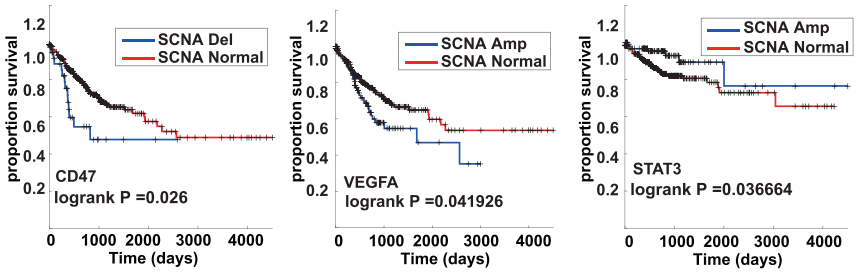

Figure 5. Kaplan-Meier Curves of Immunosuppressive DSGs

DSGs in Drug-Target Discovery

We hypothesized that if a gene was the target of a drug, its overexpression should be more sensitive to the drug; thus, we focused on the amplified DSGs. For each pair of potential drug targets, we identified the genes that were amplified in at least 10 samples, whose half maximal inhibitory concentration (IC_{50}) was significantly lower than that in the copy number of wild samples (Figure 6A). The DSGs with a p value <0.05 would then be considered as the target of the drug. At last, our work identified 22 pairs of drug targets in the CRC cell line, and a total of four DSGs (*BIRC5*, *CDK5*, *CDK4*, and *HDAC3*) and 22 drugs were found. By constructing the drug-target network (Figure 6B), the gene *BIRC5* had the highest 10 degrees, indicating that it could be the potential target of 10 drugs.

Previous research has shown that the drug IWP-2 inhibits the WNT signal,[33] and *BIRC5* is one of the essential genes in the WNT signal.[34] *CDK5* is the potential target of I-BET-762, ulixertinib, SCH772984, VX-11e, crizotinib, trametinib, and PD0325901. It was reported that the extracellular signal-regulated kinase (ERK)1/2 inhibitor SCH772984 abrogated the effects of transforming growth factor β1 (TGF-β1) on *Cdk5* and Bax levels.[35] *CDK4* can be considered as the target of AZD6738, GSK2578215A, tamoxifen, and LGK974. One study on *CDK4* and tamoxifen showed that in the cell culture, a selective *CDK4/6* inhibitor was preferentially effective in estrogen receptor-positive (ER^+) diseases and apparently acted synergistically with tamoxifen or trastuzumab,[36] and another study reported that *LEM4* overexpression renders ER^+ breast cancer cells that are resistant to tamoxifen by activating the *cyclin D-CDK4/6* axis and ERα signaling.[37] These results indicated that the drug-sensitive DSGs might be considered to be the potential target of drugs in CRC.

DISCUSSION

In this study, we constructed a new pipeline (PDSG) that can effectively identify DSGs and then systematically analyzed the effects and roles of DSGs in CRC. Compared with the existing methods, PDSG had a better performance, and the DSGs had a high consistency between solid tumors and cell lines. Additionally, the higher DSGs showed a good prognostic value for CRC patients. Our analysis also discovered that DSGs affected various aspects of CRC, such as acting as driver genes in the pathways of CRC or immune suppressor genes, in addition to their application in the drug-target relationship. Our analysis also revealed that the DSGs in MSI-H samples also have higher dosage sensitivity in MSS and MSI-L samples. Our work suggested that dosage sensitivity, which was a feature of CNA in cancer genes, played an important role in CRC, providing a new insight into CRC and a new direction for cancer treatment.

Compared with the studies about dosage effect in plant and mental retardation-related diseases, only a few studies about DSGs in cancer were reported. PDSG can relatively identify DSGs that are subject to linear and non-linear relationships between SCNA and the gene expression in an effective way. Our results revealed that there was indeed a nonlinear relationship between SCNA and the gene expression, which supplemented the previous work and provided a reference to understand this relationship. One problem with the correlation analysis using computational methods is that if one variable has a small change and the other variable has a large one, they both tend to have higher correlation scores, and this problem could make the results of the correction analysis untrustworthy. In addition, we performed differential analysis under different SCNA thresholds, mainly to screen the genes for expression changes caused by small changes in DNA SCNAs. In fact, our results also showed that our strategy was effective and comprehensive. In conclusion, PDSG has shown a good performance regarding both the pipeline design and pipeline application and may provide a reference to study DSGs in other cancers.

Our results show that the frequency of SCNA in MSI-H is lower than that in MSI-L and MSS samples, which may also be a reason for having fewer DSGs in MSI-H. The high overlap of DSGs in MSI-H samples with MSI-L and MSS samples suggests that the DSGs in MSI-H can reflect the situation in most CRC samples to a certain extent, which could help with clinical diagnosis.

The distribution analysis of DSGs in the CRC pathways showed that DSGs affected nearly every sub-pathway of CRC, indicating that the copy number amplification or deletion has an important effect on the CRC pathways. Once DSGs were amplified or deleted, the corresponding pathways produced a cascade reaction. Our study provided a foundation to further reveal the role of DSGs in cancer pathways and a new analytical direction. The drug-target analysis of DSGs suggested that the copy number amplified DSGs are less resistant to several drugs, although it was not known whether this reduction in resistance was caused by the dosage sensitivity alone, and it should be further investigated, but the results provided a new way for the precise treatment of cancer.

At present, the research about CRC has shown that SCNA has an effect on expression, but it was not clear whether this dosage effect could be observed in other cancers. There are some differences in the CNA of different cancers[38], as well as in the

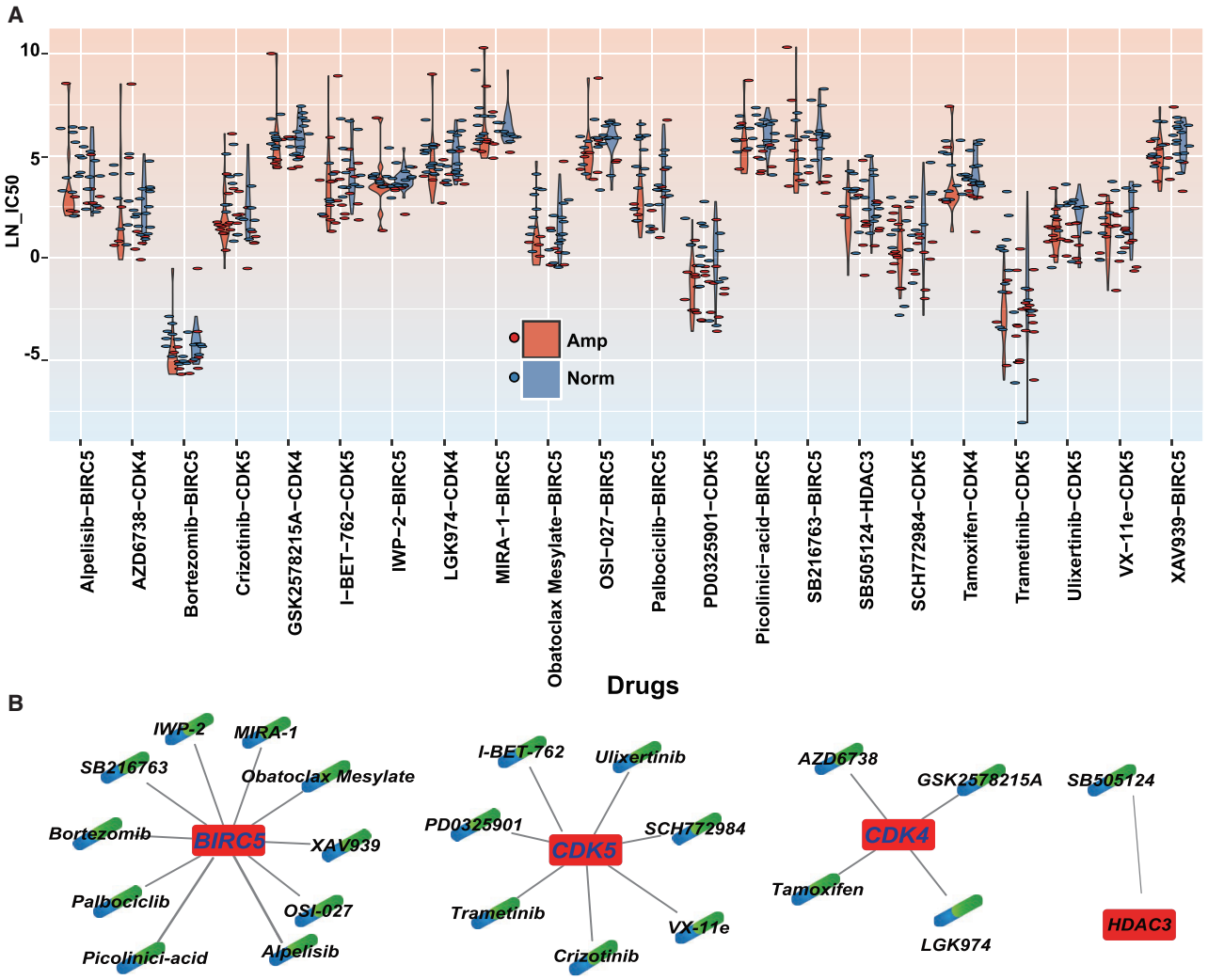

Figure 6. Application of DSGs in Drug-Target Discovery in CRC

(A) Drug-target network. The target was the amplified DSGs, and when the target was more sensitive in amplified samples than in normal samples, an edge was added to connect the drug and the target. (B) Differential analysis of ln(IC_{50}) between amplified samples and normal samples in CRC.

transcriptional regulation mechanism between genes. Therefore, more work needs to be carried out on the identification and analysis of cancer DSGs.

In summary, this study provided a good pipeline for effectively mining CRC genes that were linearly and non-linearly affected by the copy number, and systematically revealed the possible functions of CRC DSGs and their effects on CRC.

MATERIALS AND METHODS

Dataset Collection and Processing

We retrieved the level-3 RNA sequencing (RNA-seq) dataset (fragments per kilobase of transcript per million mapped reads [FPKM]) of mRNA, clinical information, and DNA copy number dataset of the Genome-Wide Human SNP Array 6.0 platform of CRC from TCGA (https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga). SCNAs were calculated by the Genomic Identification of Significant Targets in Cancer[39] (GISTIC 2.0). The genes with FPKM equal to 0 in more than 80% of the samples were filtered. Finally, we got a total of 448 CRC samples and 15,172 protein-coding genes. The cell line data of CRC, including the FPKM values of RNA-seq and CNA in the gene level, were downloaded from the CCLE (https://www.broadinstitute.org/ccle/home), and a total of 53 samples were collected. The genome variation profiling (including 114 CRC samples) and the survival information in GEO: GSE75500[40] were downloaded from Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo). The platform is Agilent-022522 SurePrint G3 comparative genomic hybridization (CGH) array 4×180K. The circular binary segmentation

algorithm identified genomic regions with abnormal copy number with DNACopy.[41]

Five Steps for Identification of DSGs

To identify the somatic copy number DSGs in CRC, we performed a five-step analysis based on the definition of DSGs:

Step 1: For each gene$_i$ (G_i) with a given SCNA value x ($x > 0$) as the threshold, the samples of CRC were divided into the following three groups: G_{i1}, copy number of amplification samples (SCNA $> x$); G_{i2}, copy number of deletion samples ($x >$ SCNA $> -x$); and G_{i3} copy number of wild samples (SCNA $< -x$). The samples were marked with $Sig(scna)$ as follows:

$$Sig(scna) = \begin{cases} 1 & if \ \ scna > x \\ 0 & if \ \ x > scna > -x \\ -1 & if \ \ scna < -x \end{cases},$$

where 1 represents amplification, 0 represents normal, and -1 represents deletion.

Step 2: Based on step 1, the Wilcoxon rank-sum test was applied to identify the differentially expressed genes between $\exp(G_{i1})$ versus $\exp(G_{i3})$ and $\exp(G_{i2})$ versus $\exp(G_{i3})$. Since a little alteration of the copy number may not cause significant changes in the expression, loose restrictions were implemented. Then, the genes were marked with the formula as follows:

$$Sig(\exp) = \begin{cases} 1 & if \ \ fold \ \ change > 1 \ \ and \ \ p < 0.05 \\ -1 & if \ \ fold \ \ change < 1 \ \ and \ \ p < 0.05 \end{cases},$$

where 1 represents upregulation and -1 represents downregulation.

Step 3: Identify the genes whose expression was consistent with its copy number. As below, if the copy number of a gene is amplification in G_{i1} and the expression is upregulation, or if the copy number of a gene is deletion in G_{i2} and the expression is downregulation, the gene is then considered to be the consistent between the copy number and gene expression as follows:

$$Sig(scna - \exp) = \begin{cases} 1 & if \ \ sig(scna) = 1 \quad and \ \ sig(\exp) = 1 \\ 1 & if \ \ sig(scna) = -1 \quad and \ \ sig(\exp) = -1 \\ 0 & others \end{cases}$$

where 1 represents that the expression is consistent with its copy number.

Step 4: Compute stable consistent scores. The SCNA threshold x was raised from 0.1 to 0.3 with 0.01 steps. The stable consistent score was computed as follows:

$$Stable\_Consistent\_Score = \bigcap_{x=0.1}^{0.3} Sig(scnv - \exp).$$

The gene with a $Stable\_Consistent\_Score = 1$ was considered to be stable in G_{i1} and G_{i2}.

Step 5: Compute dosage-sensitive scores. If a gene is copy number dosage-sensitive, it would be regulated by positive feedback. Thus, we hypothesized that the expression of one gene will have a linear or exponential change increase with the increase of SCNA. For the exponential change model, $EXP \propto r^{SCNA}$, which is equal to $\log(EXP) \propto SCNA$. The samples with $|SCNA| > 0.1$ were extracted to calculate the dosage-sensitive score, and the genes with $p < 0.01$ and $Dosage\_Sensitive\_Score > 0.5$ were considered to be DSGs.

$$Dosage\_Sensitive\_Score = \max(R_{linear}, R_{nonlinear}),$$

where

$$R_{linear} = \frac{\sum_{i=1}^{n}\left(EXP_i - \overline{EXP}\right)\left(SCNA_i - \overline{SCNA}\right)}{\sqrt{\sum_{i=1}^{n}\left(EXP_i - \overline{EXP}\right)^2}\sqrt{\sum_{i=1}^{n}\left(SCNA_i - \overline{SCNA}\right)^2}} \text{ and}$$

$$R_{nonlinear} = \frac{\sum_{i=1}^{n}\left(\log(EXP_i) - \overline{\log(EXP)}\right)\left(SCNA_i - \overline{SCNA}\right)}{\sqrt{\sum_{i=1}^{n}\left(\log(EXP_i) - \overline{\log(EXP)}\right)^2}\sqrt{\sum_{i=1}^{n}\left(SCNA_i - \overline{SCNA}\right)^2}}$$

Result Evaluation

In order to evaluate the performance of PDSG, we compared it with the linear regression pipeline that has been applied to identify DSGs in most studies, the S-curve pipeline, and the exponential transformation of the FPKM pipeline. The R value and sum of squares of residuals were calculated to evaluate the results. Moreover, the cell line dataset from CCLE was applied to evaluate the robustness of PDSG.

Perturbation Analysis

The perturbation analysis was performed to test whether the DSGs (top five DSGs in solid tumor) from the cell line have a higher dosage effect score than do the random five genes from the cell line tumor. Ten thousand perturbations were performed. In each perturbation, five genes from the cell line were randomly extracted, the average

of the dosage effect score of these five genes was calculated, and then the perturbation p value was calculated as follows:

$$Perturbation - p = \frac{1}{10000} \sum_{i=1}^{10000} avg(P_i) > x,$$

where x represents the average of dosage effect scores of the top five genes in cells, and $avg(P_i)$ represents the average of dosage effect score in the ith perturbation.

Prognosis Analysis

The progression-free interval time and disease-specific survival event data were obtained from the study of Liu et al.[42] The log-rank test and the Kaplan-Meier survival curves were used to assess the differences in survival time between G_{i1}versus G_{i3} and G_{i2} versus G_{i3}.

Construction of Drug-Target Network

IC_{50} values of CRC cell lines were retrieved from the Genomics of Drug Sensitivity in Cancer[43] (GDSC) database. The genes whose number of samples (amplified samples or wild samples) was less than 10 were filtered. For each gene-drug pair, the Wilcoxon rank-sum test was applied to calculate the significant difference of IC_{50} between the amplified and wild samples, and the gene-drug pair with a p value <0.05 and whose IC_{50} value of amplified samples was less than that in wild samples was considered to be a potential drug target. Then, the drug-target network was constructed using Cytoscape software,[44] where the nodes represent the drugs or DSGs and every edge represents a relationship between a drug and the corresponding sensitive DSGs.

Other Datasets

Sixty-one driver genes of CRC were retrieved from DriverDB v3.0 (http://driverdb.tms.cmu.edu.tw/), the CRC pathway map, and gene lists were obtained from KEGG (https://www.kegg.jp/) and 39 immunosuppressive genes were collected from the study by Zhang et al.[27] Gene function enrichment was performed using the ClusterProfiler[45] package of R language.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.omtm.2020.06.020

AUTHOR CONTRIBUTIONS

Y.X. conducted, conceived, and planned experiments; Z.C. and X.L. performed the data processing and analysis and wrote the manuscript; and W.Z. performed drug-target analysis and wrote the manuscript.

CONFLICTS OF INTEREST

The authors declare no competing interests.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grants 81673036 and 81372492) and by The Fundamental Research Funds for the Provincial Universities (31041180039).

REFERENCES

1. Siegel, R.L., Miller, K.D., and Jemal, A. (2019). Cancer statistics, 2019. CA Cancer J. Clin. *69*, 7–34.

2. Li, J., Dittmar, R.L., Xia, S., Zhang, H., Du, M., Huang, C.C., Druliner, B.R., Boardman, L., and Wang, L. (2017). Cell-free DNA copy number variations in plasma from colorectal cancer patients. Mol. Oncol. *11*, 1099–1111.

3. Oliveira, D.M., Santamaria, G., Laudanna, C., Migliozzi, S., Zoppoli, P., Quist, M., Grasso, C., Mignogna, C., Elia, L., Faniello, M.C., et al. (2018). Identification of copy number alterations in colon cancer from analysis of amplicon-based next generation sequencing data. Oncotarget *9*, 20409–20425.

4. Pongtheerat, T., and Saelee, P. (2016). Role of GSTM1 copy number variant in the prognosis of Thai colorectal cancer patients treated with 5-FU-based chemotherapy. Asian Pac. J. Cancer Prev. *17*, 4719–4722.

5. Nordick, K., and Al Khalili, Y. (2020). Genetics, Trinucleotide. In StatPearls (StatPearls Publishing).

6. Shi, J., Zhou, W., Zhu, B., Hyland, P.L., Bennett, H., Xiao, Y., Zhang, X., Burke, L.S., Song, L., Hsu, C.H., et al. (2016). Rare germline copy number variations and disease susceptibility in familial melanoma. J. Invest. Dermatol. *136*, 2436–2443.

7. Birchler, J.A., and Veitia, R.A. (2012). Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. Proc. Natl. Acad. Sci. USA *109*, 14746–14753.

8. Umehara, T., Arita, H., Yoshioka, E., Shofuda, T., Kanematsu, D., Kinoshita, M., Kodama, Y., Mano, M., Kagawa, N., Fujimoto, Y., et al. (2019). Distribution differences in prognostic copy number alteration profiles in IDH-wild-type glioblastoma cause survival discrepancies across cohorts. Acta Neuropathol. Commun. *7*, 99.

9. Mirchia, K., Sathe, A.A., Walker, J.M., Fudym, Y., Galbraith, K., Viapiano, M.S., Corona, R.J., Snuderl, M., Xing, C., Hatanpaa, K.J., and Richardson, T.E. (2019). Total copy number variation as a prognostic factor in adult astrocytoma subtypes. Acta Neuropathol. Commun. *7*, 92.

10. Smida, J., Xu, H., Zhang, Y., Baumhoer, D., Ribi, S., Kovac, M., von Luettichau, I., Bielack, S., O'Leary, V.B., Leib-Mösch, C., et al. (2017). Genome-wide analysis of somatic copy number alterations and chromosomal breakages in osteosarcoma. Int. J. Cancer *141*, 816–828.

11. Rice, A.M., and McLysaght, A. (2017). Dosage sensitivity is a major determinant of human copy number variant pathogenicity. Nat. Commun. *8*, 14366.

12. Harel, T., and Lupski, J.R. (2018). Genomic disorders 20 years on-mechanisms for clinical manifestations. Clin. Genet. *93*, 439–449.

13. Bartel, D.P. (2009). MicroRNAs: target recognition and regulatory functions. Cell *136*, 215–233.

14. Komor, M.A., de Wit, M., van den Berg, J., Martens de Kemp, S.R., Delis-van Diemen, P.M., Bolijn, A.S., Tijssen, M., Schelfhorst, T., Piersma, S.R., Chiasserini, D., et al.; NGS-ProToCol Consortium (2020). Molecular characterization of colorectal adenomas reveals POFUT1 as a candidate driver of tumor progression. Int. J. Cancer *146*, 1979–1992.

15. Liang, H., Plazonic, K.R., Chen, J., Li, W.H., and Fernández, A. (2008). Protein underwrapping causes dosage sensitivity and decreases gene duplicability. PLoS Genet. *4*, e11.

16. Maillard, A.M., Ruef, A., Pizzagalli, F., Migliavacca, E., Hippolyte, L., Adaszewski, S., Dukart, J., Ferrari, C., Conus, P., Männik, K., et al.; 16p11.2 European Consortium (2015). The 16p11.2 locus modulates brain structures common to autism, schizophrenia and obesity. Mol. Psychiatry *20*, 140–147.

17. Samur, M.K., Shah, P.K., Wang, X., Minvielle, S., Magrangeas, F., Avet-Loiseau, H., Munshi, N.C., and Li, C. (2013). The shaping and functional consequences of the dosage effect landscape in multiple myeloma. BMC Genomics *14*, 672.

18. Yan, Z., Liu, Y., Wei, Y., Zhao, N., Zhang, Q., Wu, C., Chang, Z., and Xu, Y. (2017). The functional consequences and prognostic value of dosage sensitivity in ovarian cancer. Mol. Biosyst. *13*, 380–391.

19. Veitia, R.A., Bottani, S., and Birchler, J.A. (2013). Gene dosage effects: nonlinearities, genetic interactions, and dosage compensation. Trends Genet. *29*, 385–393.

20. Clemen, C.S., Marko, M., Strucksberg, K.H., Behrens, J., Wittig, I., Gärtner, L., Winter, L., Chevessier, F., Matthias, J., Türk, M., et al. (2015). VCP and PSMF1: antagonistic regulators of proteasome activity. Biochem. Biophys. Res. Commun. *463*, 1210–1217.

21. Boland, K., Flanagan, L., McCawley, N., Pabari, R., Kay, E.W., McNamara, D.A., Murray, F., Byrne, A.T., Ramtoola, Z., Concannon, C.G., and Prehn, J.H. (2016). Targeting the 19S proteasomal subunit, Rpt4, for the treatment of colon cancer. Eur. J. Pharmacol. *780*, 53–64.

22. Borovski, T., Vellinga, T.T., Laoukili, J., Santo, E.E., Fatrai, S., van Schelven, S., Verheem, A., Marvin, D.L., Ubink, I., Borel Rinkes, I.H.M., and Kranenburg, O. (2017). Inhibition of RAF1 kinase activity restores apicobasal polarity and impairs tumour growth in human colorectal cancer. Gut *66*, 1106–1115.

23. Li, X., Stevens, P.D., Liu, J., Yang, H., Wang, W., Wang, C., Zeng, Z., Schmidt, M.D., Yang, M., Lee, E.Y., and Gao, T. (2014). PHLPP is a negative regulator of RAF1, which reduces colorectal cancer cell motility and prevents tumor progression in mice. Gastroenterology *146*, 1301–1312.e10.

24. Yang, P.H., Cheung, W.K., Peng, Y., He, M.L., Wu, G.Q., Xie, D., Jiang, B.H., Huang, Q.H., Chen, Z., Lin, M.C., and Kung, H.F. (2008). Makorin-2 is a neurogenesis inhibitor downstream of phosphatidylinositol 3-kinase/Akt (PI3K/Akt) signal. J. Biol. Chem. *283*, 8486–8495.

25. Zheng, Y., Zhou, J., and Tong, Y. (2015). Gene signatures of drug resistance predict patient survival in colorectal cancer. Pharmacogenomics J. *15*, 135–143.

26. Chung, I.F., Chen, C.Y., Su, S.C., Li, C.Y., Wu, K.J., Wang, H.W., and Cheng, W.C. (2016). DriverDBv2: a database for human cancer driver gene research. Nucleic Acids Res. *44* (D1), D975–D979.

27. Zhang, Y., Liao, G., Bai, J., Zhang, X., Xu, L., Deng, C., Yan, M., Xie, A., Luo, T., Long, Z., et al. (2019). Identifying cancer driver lncRNAs bridged by functional effectors through integrating multi-omics data in human cancers. Mol. Ther. Nucleic Acids *17*, 362–373.

28. Gao, L., Chen, K., Gao, Q., Wang, X., Sun, J., and Yang, Y.G. (2017). CD47 deficiency in tumor stroma promotes tumor progression by enhancing angiogenesis. Oncotarget *8*, 22406–22413.

29. Gao, Q., Chen, K., Gao, L., Zheng, Y., and Yang, Y.G. (2016). Thrombospondin-1 signaling through CD47 inhibits cell cycle progression and induces senescence in endothelial cells. Cell Death Dis. *7*, e2368.

30. Li, L., Liu, F., Huang, W., Wang, J., Wan, Y., Li, M., Pang, Y., and Yin, Z. (2019). Ricolinostat (ACY-1215) inhibits VEGF expression via PI3K/AKT pathway and promotes apoptosis in osteoarthritic osteoblasts. Biomed. Pharmacother. *118*, 109357.

31. Peng, N., Gao, S., Guo, X., Wang, G., Cheng, C., Li, M., and Liu, K. (2016). Silencing of VEGF inhibits human osteosarcoma angiogenesis and promotes cell apoptosis via VEGF/PI3K/AKT signaling pathway. Am. J. Transl. Res. *8*, 1005–1015.

32. Grabner, B., Schramek, D., Mueller, K.M., Moll, H.P., Svinka, J., Hoffmann, T., Bauer, E., Blaas, L., Hruschka, N., Zboray, K., et al. (2015). Disruption of STAT3 signalling promotes KRAS-induced lung tumorigenesis. Nat. Commun. *6*, 6285.

33. García-Reyes, B., Witt, L., Jansen, B., Karasu, E., Gehring, T., Leban, J., Henne-Bruns, D., Pichlo, C., Brunstein, E., Baumann, U., et al. (2018). Discovery of inhibitor of Wnt production 2 (IWP-2) and related compounds as selective ATP-competitive inhibitors of casein kinase 1 (CK1) δ/ε. J. Med. Chem. *61*, 4087–4102.

34. Dey, N., Young, B., Abramovitz, M., Bouzyk, M., Barwick, B., De, P., and Leyland-Jones, B. (2013). Differential activation of Wnt-β-catenin pathway in triple negative breast cancer increases MMP7 in a PTEN dependent manner. PLoS ONE *8*, e77425.

35. Zhao, W., Yan, J., Gao, L., Zhao, J., Zhao, C., Gao, C., Luo, X., and Zhu, X. (2017). Cdk5 is required for the neuroprotective effect of transforming growth factor-β1 against cerebral ischemia-reperfusion. Biochem. Biophys. Res. Commun. *485*, 775–781.

36. Sutherland, R.L., and Musgrove, E.A. (2009). CDK inhibitors as potential breast cancer therapeutics: new evidence for enhanced efficacy in ER^+ disease. Breast Cancer Res. *11*, 112.

37. Gao, A., Sun, T., Ma, G., Cao, J., Hu, Q., Chen, L., Wang, Y., Wang, Q., Sun, J., Wu, R., et al. (2018). LEM4 confers tamoxifen resistance to breast cancer cells by activating cyclin D-CDK4/6-Rb and ERα pathway. Nat. Commun. *9*, 4180.

38. Cooper, L.A., Demicco, E.G., Saltz, J.H., Powell, R.T., Rao, A., and Lazar, A.J. (2018). PanCancer insights from The Cancer Genome Atlas: the pathologist's perspective. J. Pathol. *244*, 512–524.

39. Mermel, C.H., Schumacher, S.E., Hill, B., Meyerson, M.L., Beroukhim, R., and Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. *12*, R41.

40. van den Broek, E., Krijgsman, O., Sie, D., Tijssen, M., Mongera, S., van de Wiel, M.A., Belt, E.J., den Uil, S.H., Bril, H., Stockmann, H.B., et al. (2016). Genomic profiling of stage II and III colon cancers reveals APC mutations to be associated with survival in stage III colon cancer patients. Oncotarget *7*, 73876–73887.

41. Venkatraman, E.S., and Olshen, A.B. (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics *23*, 657–663.

42. Liu, J., Lichtenberg, T., Hoadley, K.A., Poisson, L.M., Lazar, A.J., Cherniack, A.D., Kovatich, A.J., Benz, C.C., Levine, D.A., Lee, A.V., et al. (2018). An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. Cell *173*, 400–416.e11.

43. Yang, W., Soares, J., Greninger, P., Edelman, E.J., Lightfoot, H., Forbes, S., Bindal, N., Beare, D., Smith, J.A., Thompson, I.R., et al. (2013). Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. *41*, D955–D961.

44. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., 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.

45. Yu, G., Wang, L.G., Han, Y., and He, Q.Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS *16*, 284–287.