JNCI Cancer Spectrum (2019) 3(2): pkz012

doi: 10.1093/jncics/pkz012
First published online April 16, 2019
Article

ARTICLE

Scarcity of Recurrent Regulatory Driver Mutations in Colorectal Cancer Revealed by Targeted Deep Sequencing

Rebecca C. Poulos, Dilmi Perera, Deborah Packham, Anushi Shah, Caroline Janitz, John E. Pimanda, Nicholas Hawkins, Robyn L. Ward, Luke B. Hesson, Jason W. H. Wong

See the Notes section for the full list of authors' affiliations.
Correspondence to: Jason W. H. Wong, School of Biomedical Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, L1, Laboratory Block, 21 Sassoon Rd, Pokfulam, Hong Kong (e-mail: jwhwong@hku.hk).

Abstract

Background: Genetic testing of cancer samples primarily focuses on protein-coding regions, despite most mutations arising in noncoding DNA. Noncoding mutations can be pathogenic if they disrupt gene regulation, but the benefits of assessing promoter mutations in driver genes by panel testing has not yet been established. This is especially the case in colorectal cancer, for which few putative driver variants at regulatory elements have been reported.
Methods: We designed a unique target capture sequencing panel of 39 colorectal cancer driver genes and their promoters, together with more than 35 megabases of regulatory elements focusing on gene promoters. Using this panel, we sequenced 95 colorectal cancer and matched normal samples at high depth, averaging 170× and 82× coverage, respectively.
Results: Our target capture sequencing design enabled improved coverage and variant detection across captured regions. We found cases with hereditary defects in mismatch and base excision repair due to deleterious germline coding variants, and we identified mutational spectra consistent with these repair deficiencies. Focusing on gene promoters and other regulatory regions, we found little evidence for base or region-specific recurrence of functional somatic mutations. Promoter elements, including *TERT*, harbored few mutations, with none showing strong functional evidence. Recurrent regulatory mutations were rare in our sequenced regions in colorectal cancer, though we highlight some candidate mutations for future functional studies.
Conclusions: Our study supports recent findings that regulatory driver mutations are rare in many cancer types and suggests that the inclusion of promoter regions into cancer panel testing is currently likely to have limited clinical utility in colorectal cancer.

In recent years, hundreds of novel cancer driver genes have been characterized through analyses made possible by the completion of large-scale cancer genome-sequencing projects. Such genes have been classified as cancer drivers because they harbor frequent high-impact somatic coding mutations in cancer genomes. Identifying cancer driver mutations outside of protein-coding elements, however, has proven to be a complex task because it can be difficult to assign function to some noncoding mutations (1). Despite a number of large-scale studies aimed at prioritizing either recurrent or functional mutations (2–4), relatively few somatic driver mutations have been discovered in the noncoding genome. One reason for the apparent sparsity of noncoding drivers may be that datasets are currently underpowered to detect mutations at low to moderate frequencies from among the considerable background of somatic passenger mutations in the cancer genome (5,6).

Received: September 26, 2018; **Revised:** February 7, 2019; **Accepted:** February 22, 2019

Detection of driver mutations in colorectal cancer has been shown to be an effective means of enabling therapy tailored to an individual patient. For example, colorectal cancer patients with *RAS* mutations respond poorly to anti–epidermal growth factor receptor cetuximab and panitumumab, and screening of patients for *RAS* mutations is now a routine aid in the decision of whether to administer anti–epidermal growth factor receptor therapy (7). Currently, targeted sequencing of specific cancer driver genes is typically achieved either by target enrichment through capture probes or by amplicon sequencing. Thus far, these methods have generally been designed to examine protein coding regions. Determining whether mutations exist in regulatory elements of cancer driver genes has not been examined in colorectal cancer beyond the use of whole-genome sequencing (WGS) datasets, which are of relatively low coverage (40×) (8). Recently, sequencing data from an assay capturing regulatory elements in addition to protein-coding regions in a large cohort of breast cancers led to the identification of recurrent somatic mutations in the promoter of the cancer driver gene *FOXA1* (5). The results of this study suggest that a targeted sequencing approach that includes regulatory regions may be useful for identifying regulatory driver mutations in other cancer types.

In our study, we performed target capture sequencing (TCS) to generate high-depth sequencing data across the promoter elements and coding regions of 39 colorectal cancer driver genes from 95 colorectal cancer and matched normal samples. In addition, we incorporated more than 35 megabases of selected regulatory regions, focusing on the promoters of all coding genes, and we comprehensively assessed somatic mutations in regulatory elements in search of noncoding driver mutations.

Methods

Patient Samples

A total of 95 colorectal cancer and matched normal samples were selected from a biobank of fresh tumor tissue and blood collected with patient informed consent. The study was carried out with ethics approval from the South Eastern Sydney Local Health District Human Research Ethics Committee (approval number H00/022 and 00113), and all samples were collected with written informed consent from each subject. Patient and sample characteristics can be found in Supplementary Tables 1 and 2 (available online).

TCS Assay Design and Analysis of Sequencing Data

A unique TCS assay encompassing 35 726 928 nucleotides of the genome was designed to provide sequencing data covering promoters, other regulatory regions, and some coding exons. WGS was also performed on one sample (see Supplementary text, available online). See Mendeley data for description and list of regions and details of analysis.

Variant Detection and Analysis

Somatic single-nucleotide variants were detected with Strelka (9). We defined high-confidence somatic mutations as those with variant allele frequencies 8.5% or higher (see Supplementary Figure 1A, available online) because this threshold led to sample-specific TCS mutation loads that were similar to those of other colorectal cancers (10), indicating mutations that are more likely to be true positives. Somatic indels were detected using Strelka (9) as well as SvABA (11) and Lancet (12) (see Supplementary text [available online] for detailed description of analysis).

Mutational Spectrum Analyses

Mutational spectra were assessed through Pearson correlation of trinucleotide frequencies in a given sample with mutational signatures (13) from the Catalogue of Somatic Mutations in Cancer (COSMIC) (14) database. Mutational spectra from TCS were normalized using genome trinucleotide frequencies. All Pearson correlations (r) reported had P less than .0001, indicating a correlation coefficient that is statistically significantly different from zero.

Experimental Validation of Variants Detected

Some somatic mutations were randomly selected for experimental validation via Sanger sequencing of polymerase chain reaction product amplified from cancer and matched normal patient DNA. Indels in the putative promoter of *MTERFD3* were also validated using the same method.

Data Availability

TCS and WGS data from this study are available upon application via the European Genome-Phenome Archive with accession number EGAD00001004582. Mendeley data associated with this study can be downloaded from http://dx.doi.org/10.17632/c65rsd4fr2.1.

Results

TCS and Mutation Detection

We designed a TCS assay encompassing 35 726 928 nucleotides of the genome (Figure 1A). The assay was designed to focus on the promoters and exons of a panel of known colorectal cancer-associated genes (n = 39 genes; see Mendeley data). The assay also incorporated promoter regions of all known coding genes (n = 26 455 regions) and selected regulatory elements, including DNase I hypersensitivity sites (n = 13 891 regions), long noncoding RNAs (n = 842 regions), and microRNAs (n = 25 regions). With this unique TCS assay, we sequenced 95 colorectal cancer and matched normal samples (Supplementary Tables 1 and 2, available online). Average (standard deviation [SD]) reads per sequenced base were 169.96 (25.08) (SD calculated across 95 samples) in the cancer samples and 81.81 (17.13) in the matched normal samples. (Figure 1B). The percentage on-target rate across TCS cancer samples was 82.33 (1.72)% (see Mendeley data).

We identified 43 915 somatic single-nucleotide mutations across our cohort, with a median of 178 mutations per sample (see Methods; Supplementary Table 2, available online) and identified 5244 somatic indels across our cohort, with a median of seven indels per sample (see Methods; Supplementary Figure 1B, Table 2, available online). We validated a selection of point mutations and a small deletion via Sanger sequencing (see Methods; Supplementary Figure 1, C and D, available online). To ensure that our TCS assay was able to capture somatic mutations in regulatory regions, we additionally performed WGS on sample

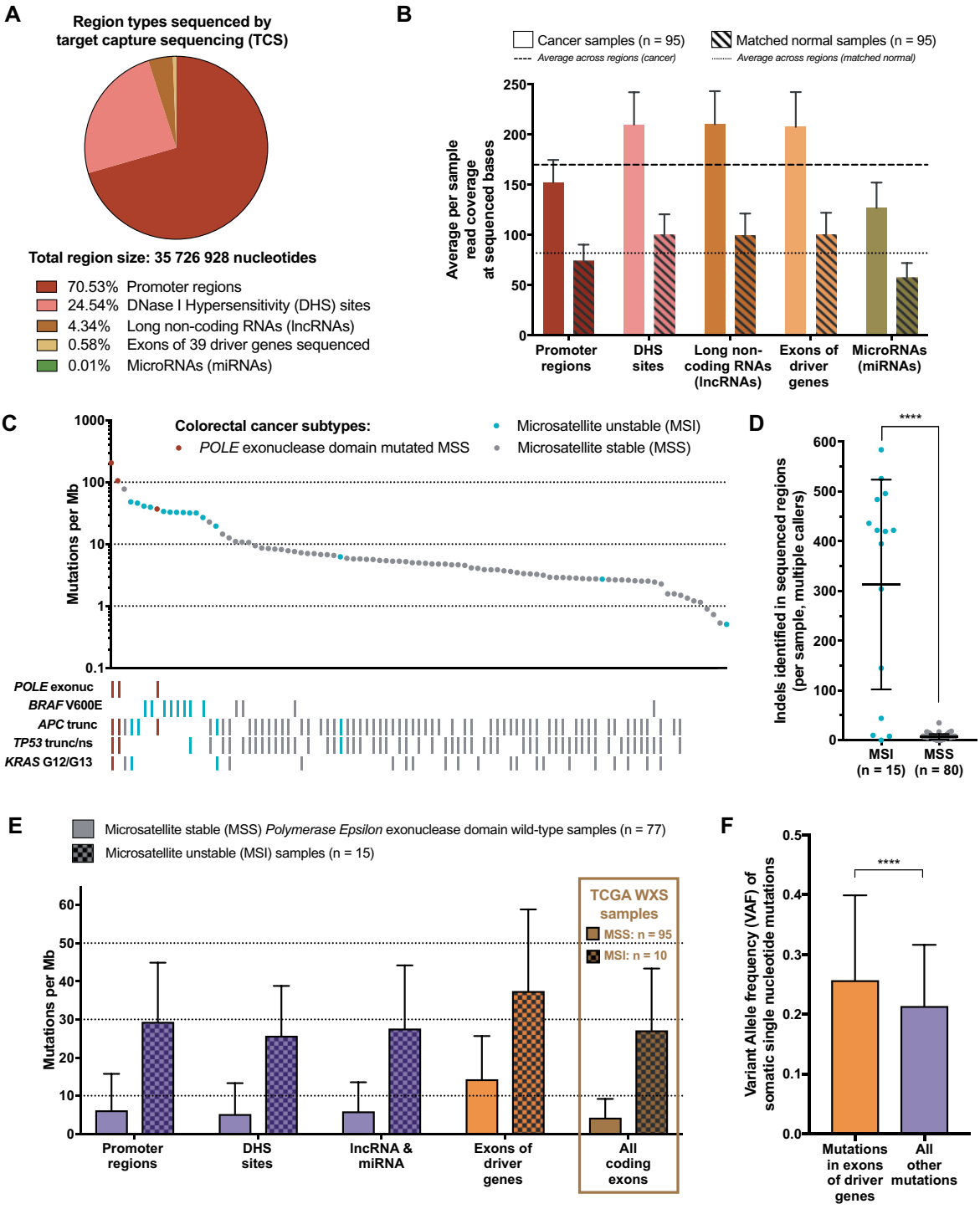

Figure 1. Sequence coverage by target capture sequencing (TCS) and mutation characteristics. **A**) Types and sizes of regions sequenced by TCS. **B**) Average per sample read coverage across sequenced bases in cancer and matched normal TCS samples. Read coverage is plotted for each region type. **Dotted lines** indicate average read coverage in cancer and matched normal samples across the cohort. **C**) Somatic single-nucleotide mutation rate per megabase (Mb) of each sample in the TCS cohort (n = 95), plotted on a log scale (*y*-axis). **Colors** represent colorectal cancer subtypes as indicated, and somatic single-nucleotide and indel mutations in colorectal cancer driver genes of interest are marked by **bars**. *POLE Exonuc* = *Polymerase Epsilon* exonuclease domain mutation; trunc = frameshift or stop-gain mutation; *ns* = nonsynonymous mutation. **D**) Numbers of indels identified in microsatellite unstable (MSI) and microsatellite stable (MSS) colorectal cancer samples sequenced by TCS. Individual samples are indicated by **dots** corresponding to the number of indels identified by at least two variant detectors. **E**) Average somatic single-nucleotide mutation rate (mutations per Mb) in cancer samples across region types sequenced by TCS. Data from somatic single-nucleotide mutations in The Cancer Genome Atlas (TCGA) whole-exome sequenced (WXS) colorectal cancer samples are shown, as indicated on the rightmost **bars** on the graph. WXS samples harboring nonsynonymous *POLE* coding variants are excluded, and coding exons are here derived from GENCODE (35) v29 data. **F**) Average variant allele frequency (VAF) of somatic single-nucleotide mutations overlapping exons of sequenced driver genes and VAF of all other mutations. In all plots, mid-line and error bars indicate mean and standard deviation, and ****$P < .0001$.

CRC_1, which had the highest mutation load. Over the regions captured by TCS, we found that the TCS assay identified 94.5% (n = 4585 of 4854) of mutations detected by WGS, suggesting that the assay is robust (for detailed comparison, see Supplementary text and Supplementary Figures 2–5, available online).

Characterization of Cohort According to Mutation Profiles

Mutation loads across our cohort were mostly consistent with previous observations among colorectal cancers (10,15) (Figure 1C), with generally increasing numbers of mutations across samples that were microsatellite stable (MSS; n = 80) and then microsatellite unstable (MSI; n = 15). We found that three MSS samples with high mutation loads harbored *POLE* exonuclease domain mutations (CRC_1: p.Pro286Arg; CRC_2: p.Met444Lys; CRC_8: p.Ser297Phe), which is known to result in an ultramutator phenotype (see Supplementary text, available online, for description of MSI and *POLE* mutation status annotation). As expected, MSI samples harbored statistically significantly more indels compared with MSS samples (Figure 1D). MSI samples in our cohort more commonly harbored *BRAF* p.Val600Glu (V600E) mutations (MSI: 8 of 15; $P < .0001$, Fisher's exact test; Figure 1C) and *RNF43* mutations (MSI: 7 of 15; $P < .0001$, Fisher's exact test), and less commonly harbored *APC* variants (MSI: 5 of 15; $P = .0044$, Fisher exact test), than did the *POLE* exonuclease domain wild-type MSS samples (*BRAF V600E* mutation: 4 of 77; *RNF43* mutation: 1 of 77; *APC* mutation: 58 of 77). We found recurrent mutations in genes such as *SMAD4* (n = 10/95), *ARID1A* (n = 8/95), and *SOX9* (n = 8/95) (Table 1), which have also previously been reported in colorectal cancer (10). Mutation loads and variant allele frequencies of somatic single-nucleotide mutations across different types of genomic regions captured in our TCS assay are shown in Figure 1, E and F, respectively.

Association of Deleterious Germline Mutations With Mutational Signatures

We next analyzed germline variants in our samples, finding three patients with putative pathogenic germline variants using ClinVar (16) annotations (see Methods and Mendeley data). These variants are nonsynonymous and truncating heterozygous variants in regions coding for *MSH6*, *MUTYH*, and *ATM*. The two samples that harbored a variant in either *MSH6* or *MUTYH* showed distinctive mutational spectra based on our TCS assay, and we examined these variants further.

Sample CRC_4 harbored a germline heterozygous C > T single-nucleotide polymorphism at chr2: 48 030 588 (Supplementary Figure 6A, available online; see Mendeley data), resulting in the introduction of an early stop codon at p.Arg1068Ter, a variant recorded in the InSiGHT database (17) as Class 5 pathogenic. A potential somatic second hit is a truncating G > A mutation in *MSH6* at chr2: 48 026 216 (p.Trp365Ter). Sample CRC_4 was the fourth-most highly mutated sample in our cohort, and we found indel mutation numbers in this sample to be elevated when compared with MSS samples. The mutational spectrum in CRC_4 was strongly correlated with mismatch repair deficiency-associated signatures 14 and 6 (13, 18) from the COSMIC (14) database ($r = 0.784$ and $r = 0.767$, respectively, with $P < .0001$ by Pearson correlation; signature 14 shown in Figure 2A). Together with the relatively early age of colorectal cancer diagnosis in this patient (51 years, presenting with synchronous cancers of the rectum and sigmoid) and the microsatellite instability we observed in sample CRC_4, these findings are consistent with Lynch syndrome.

Table 1. Summary of driver mutations in genes sequenced, with more than 5% recurrence in the sequenced cohort*

Gene	MSS (n = 77)	MSI (n = 15)	*POLE* mut. (n = 3)	Total (n = 95)
APC	58 (75.3%)	5 (33.3%)	3 (100.0%)	66 (69.5%)
TP53	53 (68.8%)	2 (13.3%)	2 (66.7%)	57 (60.0%)
KRAS	22 (28.6%)	2 (13.3%)	1 (33.3%)	25 (26.3%)
FBXW7	11 (14.3%)	4 (26.7%)	2 (66.7%)	17 (17.9%)
PIK3CA	10 (13.0%)	2 (13.3%)	3 (100.0%)	15 (15.8%)
BRAF	6 (7.8%)	8 (53.3%)	0 (0.0%)	14 (14.7%)
TCF7L2	7 (9.1%)	1 (6.7%)	2 (66.7%)	10 (10.5%)
SMAD4	8 (10.4%)	2 (13.3%)	0 (0.0%)	10 (10.5%)
RNF43	1 (1.3%)	7 (46.7%)	1 (33.3%)	9 (9.5%)
SOX9	7 (9.1%)	1 (6.7%)	0 (0.0%)	8 (8.4%)
ARID1A	2 (2.6%)	4 (26.7%)	2 (66.7%)	8 (8.4%)
PTPRK	2 (2.6%)	4 (26.7%)	1 (33.3%)	7 (7.4%)
PMS1	3 (3.9%)	1 (6.7%)	1 (33.3%)	5 (5.3%)
MSH6	1 (1.3%)	2 (13.3%)	2 (66.7%)	5 (5.3%)

*The number of samples with at least one somatic single-nucleotide or indel mutation in each gene is shown. For a mutation to be considered a driver, it must be either a frameshift or stop-gain mutation, or a missense mutation with a PROVEAN (37) converted rankscore of more than 0.5, a PolyPhen (38) prediction of "deleterious," or a SIFT (39) prediction of "possibly damaging" or "probably damaging." The percentage listed indicates the fraction of the subtype or total cohort that harbored at least one mutation fulfilling these criteria. MSI = microsatellite unstable sample; MSS = microsatellite stable sample; *POLE* mut. = Polymerase epsilon exonuclease domain-mutated MSS sample.

Our analysis of germline variants also led to the detection of one sample with a heterozygous germline C > T variant in the *MUTYH* gene at chr1: 45 798 117 (Supplementary Figure 6B; see Mendeley data). This variant has an allele frequency of 1.339×10^{-4} in the Exome Aggregation Consortium database (19), and it results in a nonsynonymous amino acid change in *MUTYH* (p.Arg242His). The variant has been shown in vitro to severely impair glycosylase and DNA binding activity (20). The mutation burden of MSS sample CRC_3 is unusual (n = 2767 mutations) because it is higher than all MSI samples in our cohort. We did not observe a putative second somatic hit in *MUTYH*, and the region did not exhibit signs of loss of heterozygosity. However, interestingly the mutational spectrum of this sample was highly correlated with the COSMIC (14) database's signature 18 ($r = 0.825$ and $P < .0001$ by Pearson correlation; Figure 2B), and this signature has been associated with defects in the base excision repair pathway and *MUTYH* deficiency (21).

Mutations in the Promoters of Cancer Driver Genes

Having shown that our TCS method accurately captured mutational spectra and both germline and somatic variants, we then examined mutations in the promoters of our set of 39 driver genes to search for putative recurrent regulatory driver mutations. We identified 33 somatic single-nucleotide and indel mutations in the promoters of these genes (see Mendeley data). Of the 39 promoters, *TCF7L2*, *VTI1A*, *SOX9*, *BMP3*, *APC*, *RSPO3*, and *BRAF* were mutated in two or more samples. To assess the potential impact of these mutations, we sought to identify a second hit to each of these genes in the same sample, such as a deleterious coding mutation or a loss of heterozygosity event (where this was possible to determine using heterozygous

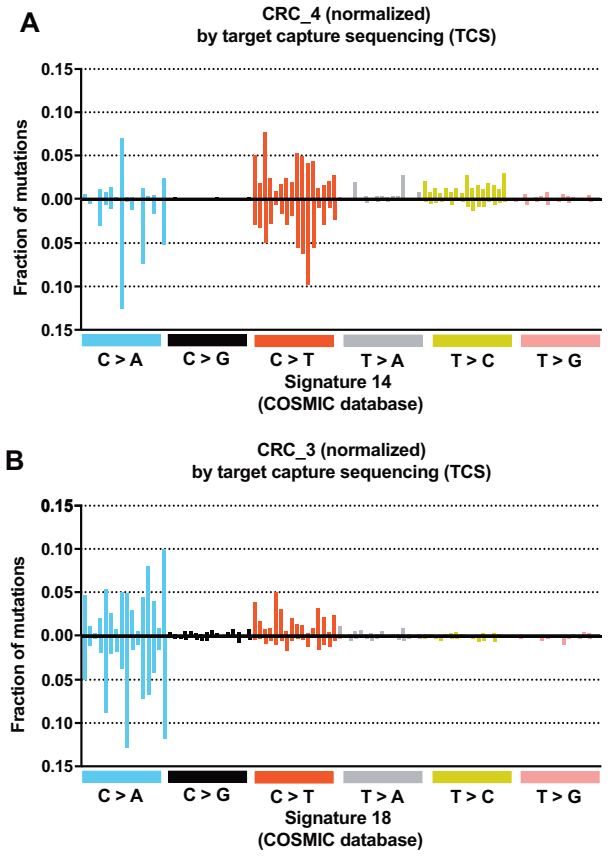

Figure 2. Mutational spectra from target capture sequencing (TCS) samples with notable germline variants. **A**) Normalized mutational spectrum from colorectal cancer sample CRC_4 (**upper**) against signature 14 from the Catalogue of Somatic Mutations in Cancer (COSMIC) (14) database (**lower**). **B**) Normalized mutational spectrum from colorectal cancer sample CRC_3 (**upper**) against signature 18 from the COSMIC (14) database (**lower**).

germline single-nucleotide polymorphisms). We found one promoter mutation each in *SOX9* and *APC* that co-occurred with a corresponding loss-of-function mutation event. However, we found loss-of-function mutations to be common in these genes across all samples, and it is therefore not possible to conclude whether these promoter mutations are clear driver events. To further assess whether any of the seven recurrently mutated promoters warranted further investigation, we considered mutations within these promoters in an additional colorectal cancer cohort from The Cancer Genome Atlas (TCGA) (n = 60 samples analyzed; see Mendeley data). Interestingly, we found that seven TCGA samples also harbored a mutation in the promoter of *BMP3* (Figure 3A; see Mendeley data). Despite this, analysis of sample-matched gene expression data from TCGA demonstrated no statistically significant difference between *BMP3* promoter wild-type and mutated samples ($P = .9809$ by unpaired *t* test; Figure 3B). The only other genes that were recurrently mutated at the promoter both in the TCS and TCGA cohorts were *APC* and *VTI1A* (see Mendeley data), but expression was not statistically significantly different between promoter wild-type and mutated samples (*APC*: $P = .7033$ and *VTI1A*: $P = .3704$, respectively; Figure 3B). Taken together, across the promoters of the 39 cancer driver genes sequenced, we found that functional promoter driver mutations are either absent or rare events in colorectal cancer.

We also captured the promoter of the *TERT* gene in our TCS assay, a gene that harbors the most frequent regulatory driver mutations yet discovered in cancer (22,23). We investigated whether samples in our cohort also harbored somatic mutations at chr5: 1, 295, 228 and chr5: 1, 295, 250 within the *TERT* promoter, but we found that none of our cancer samples carried these mutations (see Mendeley data). The *TERT* promoter was also devoid of any somatic single-nucleotide or indel mutations, confirming previous reports that such mutations are of low frequency in colorectal cancer (24,25).

Search for Other Potentially Functional Mutations in Regulatory Regions

We next utilized our full mutation dataset covering all captured regions in search of other potentially functional regulatory mutations. We used OncodriveFML to search for functional enrichment in genomic regions (26). We did not find any functional enrichment of mutations in our cohort beyond coding exons of the driver genes sequenced (Figure 3C). Assigning function to a noncoding variant can be imprecise because of the variety of ways in which a variant may affect gene regulation (1), so we also considered base pair recurrence of somatic variants within our cohort. To improve filtering of variants in our TCS cohort (n = 95 samples), we also incorporated single-nucleotide variants from WGS colorectal cancer samples from TCGA (n = 60; see Mendeley data). Within regulatory regions, we found 90 recurrent somatic single-nucleotide variants that were present in more than three samples, with at least one sample from each of the TCS and TCGA cohorts (see Mendeley data). These variants were then prioritized by FunSeq2 (27), which selected 47 variants (see Mendeley data) as candidate functional mutations by designating a high noncoding variant score or an association with any cancer gene. These mutations were in proximity to cancer-related genes such as *JUN*, *CDKN1B*, and *ASF1A*. Further investigating these 47 mutations identified as candidates by FunSeq2, we selected any mutations that were present in four or more TCGA colorectal cancers and examined expression of nearby gene(s) in wild-type and mutant samples (n = 4 mutations selected). We investigated the expression of any gene within 6 kilobases (kb) of the variant but found no statistically significant difference in gene expression via RNA-sequencing by unpaired *t* test between TCGA samples that were wild-type and mutant for each variant.

Finally, we investigated recurrent indel mutations that lay outside of the coding exons of the driver genes sequenced. Counting indels in our TCS cohort that arose within our sequenced regions in defined genomic windows (indel locus ± 10 base pairs [bp]), we identified 16 candidate windows (21–28 bp) harboring a total of 66 indels (see Methods and Mendeley data). Windows were selected as candidates if they harbored at least four indels, or three or more indels if at least one sample was MSS. With the exception of the recurrent indels within the *VTI1A* promoter discussed previously, the noncoding indel with the highest transcription factor-occupancy lay within the region chr12: 107, 380, 956–107, 380, 983 (indels overlapping a maximum of 46 transcription factor chromatin immunoprecipitation sequencing annotations; n = 3 indels). The region lies within a putative promoter for the mitochondrial transcription termination factor (mTERF) *MTERFD3* (Figure 3D), and we validated all three indels via Sanger sequencing (Figure 4A). The indels in our cohort overlap Factorbook (28) binding sites for transcription factors *SP1/SP2*, *E2F4/E2F6*, and *MAZ* (Figure 4B). Overexpression

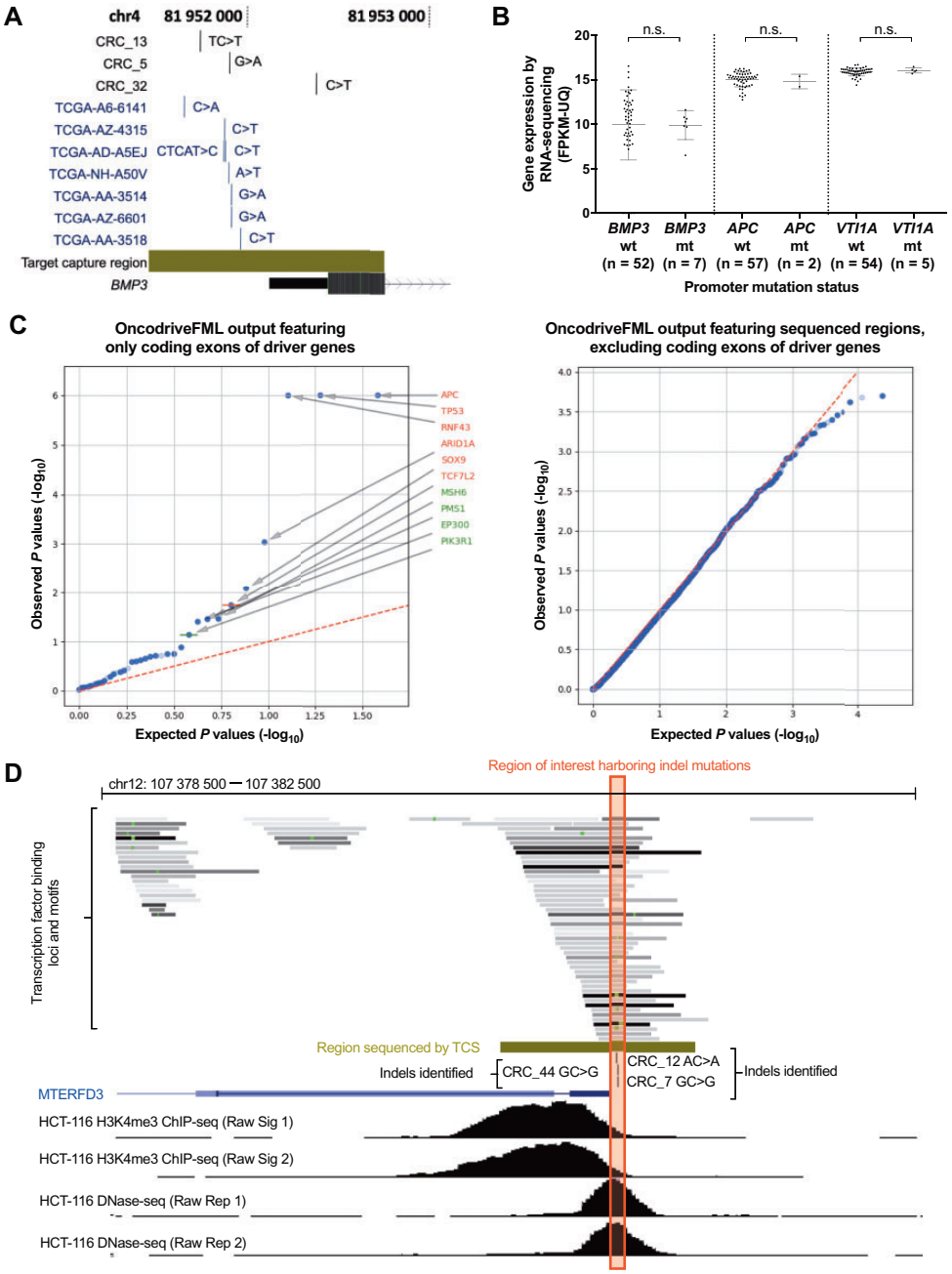

Figure 3. Search for putative driver variants in target capture sequencing (TCS) data. **A)** Snapshot from the University of California Santa Cruz (UCSC) Genome Browser, indicating the location of somatic mutations from our TCS and The Cancer Genome Atlas (TCGA) colorectal cancer cohort within the promoter of *BMP3*. **B)** Expression of *BMP3* (**left**), *APC* (**middle**), and *VTI1A* (**right**) in promoter wild-type (wt) and mutant (mt) TCGA colorectal cancer samples, for the respective genes. n.s = not statistically significant by unpaired *t* test; mean and standard deviation are shown. **C)** Quantile-quantile plots produced by OncodriveFML (26) showing the expected and observed distribution of *P* values demonstrating any functional somatic variant bias in coding exons of the colorectal cancer-associated genes sequenced (**left**) and all sequenced regions, excluding coding exons from sequenced colorectal cancer-associated genes (**right**). **Dots** represent different sequenced regions, with **lighter colors** indicating regions for which the number of mutated samples did not reach the minimum required to perform the multiple testing correction. Sequenced regions identified as statistically significant are indicated for q-value less than 0.1 (**red**) and q-value less than 0.25 (**green**). **D)** Snapshot from UCSC Genome Browser, indicating the location of indels within the putative promoter of *MTERFD3*. Transcription factor binding data are shown from the ENCODE (36) "Transcription Factor ChIP-seq (161 factors)" track. **Grey boxes** indicate peak clusters of transcription factor occupancy, where the **darkness** of each box signifies the maximum signal strength observed in any cell line contributing to that cluster. A **green highlight** within a box designates the site of the highest scoring canonical motif for the transcription factor indicated, via Factorbook (28) annotations. HCT-116 (human colon cancer cell-line) H3K4me3 chromatin immunoprecipitation sequencing (ChIP-seq) and DNase I hypersensitivity sequencing (DNase-seq) data are also shown.

of *MTERFD3* and other mTERF family proteins has been associated with mitochondrial DNA copy number depletion (29), and mitochondrial DNA copy number variation can occur in cancer tissues (30). However, experimental functional validation would be required to determine whether these variants might contribute toward oncogenesis through such a capacity.

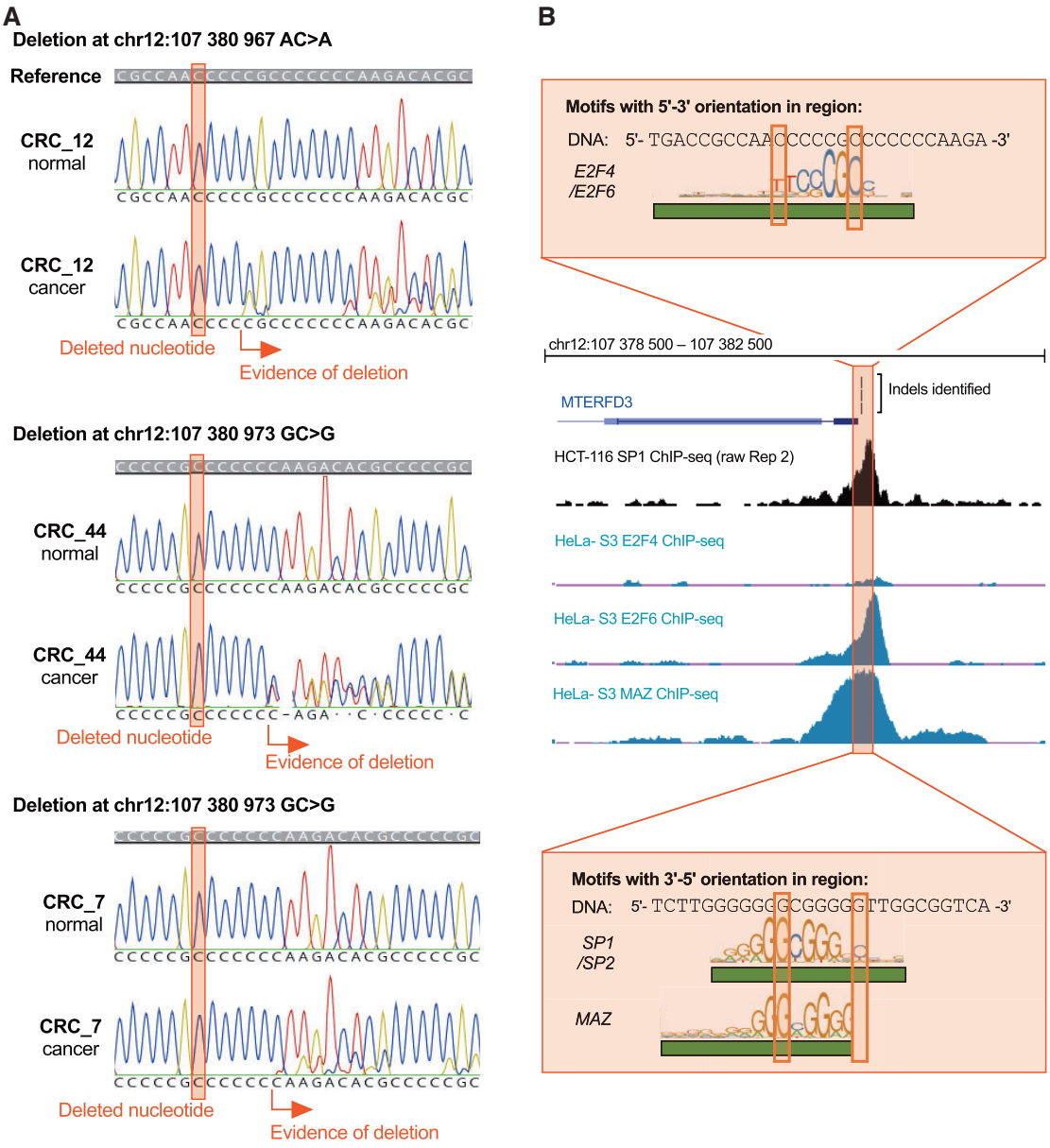

Figure 4. Validation by Sanger sequencing, and the genomic locus harboring deletions in the *MTERFD3* putative promoter. **A)** Sequencing traces from Sanger sequencing of genomic DNA, depicting validation of the three indels within the *MTERFD3* putative promoter. Sequencing traces are visualized using Geneious version 10.2.2 (http://www.geneious.com). **B)** Snapshot from the University of California Santa Cruz Genome Browser, indicating deletions (indels) within the putative promoter of *MTERFD3*, alongside chromatin immunoprecipitation sequencing (ChIP-seq) data for the transcription factors with motifs disrupted. **Boxes** contain the reference DNA sequence, with the deleted nucleotides marked by an **orange box**. Transcription factor binding motifs are shown from Factorbook (28), where a **green bar** depicts the span of the motif across the DNA sequence.

Discussion

In recent years, many recurrent mutations have been found within *cis*-regulatory regions of cancer genomes, but few drivers have yet been identified. We undertook this study to detect regulatory driver mutations in colorectal cancer and to determine whether it may be beneficial to include the sequencing of gene promoters into current somatic mutation testing panels in colorectal cancer. Our results suggest there are no strong candidate recurrent regulatory driver mutations in the promoters of key colorectal cancer driver genes and other regulatory regions that we sequenced. Our findings suggest such mutations may be rare in colorectal cancer more generally. Because of the current sparsity of recurrent regulatory driver mutation discoveries in colorectal cancer, it is unlikely to be beneficial at this stage to include the sequencing of gene promoters in current panel testing assays.

In this study, we list single-nucleotide variants and genomic windows containing recurrent indels that may be functional noncoding mutations. We selected these variants by measuring recurrence, FunSeq2 score (27), and annotations of transcription factor binding. There are a plethora of ways in which regulatory mutations may affect genome function. For example, a mutation may alter a transcription factor binding site, affect the partitioning of the genome into topologically associating domains, or cause epigenetic changes by altering the binding of pioneer

factors, nucleosome positioning, chromatin organization, or CpG methylation (1). We have ultimately based our functional annotations on computational prediction methods, and it is possible that recurrent mutations that we identified affect gene regulation in a way that is not captured by these prediction methods. Finally, in some instances, site-specific sequence contexts may limit sequencing coverage across particular regulatory elements and limit the detection of certain driver variants, as described (5).

This study can serve as a reference point and validation cohort for future work that could utilize TCS to further increase cohort sizes in cancer and to enable the detection of low-frequency regulatory driver mutations. The variants that we identified may also be reanalyzed as newer computational approaches are developed for the identification of functional regulatory mutations. Our TCS approach could also be effectively applied to noncoding driver detection in melanoma and blood cancers, for which a number of somatic regulatory driver mutations have already been well established (23,31–34).

Funding

This work was supported by Cancer Institute New South Wales (13/DATA/1-02, JWHW); Cure Cancer Foundation Australia with the assistance of Cancer Australia, through the Priority-driven Collaborative Cancer Research Scheme (APP1057921, JWHW); Australian Research Council Future Fellowship (FT130100096, JWHW); Australian Government Research Training Program Scholarship (RCP); and National Health and Medical Research Council (Australia) (APP1138536, RCP) (JEP).

Notes

Affiliations of authors: Prince of Wales Clinical School and Lowy Cancer Research Centre, UNSW Sydney, Sydney, NSW, Australia (RCP, DPe, DPa, AS, JEP, RLW, LBH, JWHW); Children's Medical Research Institute, Faculty of Medicine and Health The University of Sydney, Westmead, NSW, Australia (RCP); Next-Generation Sequencing Facility, Office of the Deputy Vice-Chancellor (R&D), Western Sydney University, Penrith, NSW, Australia (CJ); Department of Haematology, Prince of Wales Hospital, Sydney, NSW, Australia (JEP); School of Medical Sciences, UNSW Sydney, Sydney, NSW, Australia (JEP, NH); Faculty of Medicine, The University of Queensland, Herston, QLD, Australia (NH); Faculty of Medicine and Health, University of Sydney, Sydney, NSW, Australia (RLW); School of Biomedical Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong Special Administrative Region (JWHW).

References

1. Poulos RC, Wong J. Cis-regulatory driver mutations in cancer genomes. In: *eLS*. Chichester: John Wiley & Sons, Ltd; 2017:1–10.
2. Melton C, Reuter JA, Spacek DV, et al. Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nat Genet*. 2015;47(7):710–716.
3. Fredriksson NJ, Ny L, Nilsson JA, et al. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nat Genet*. 2014;46(12):1258–1263.
4. Weinhold N, Jacobsen A, Schultz N, et al. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat Genet*. 2014;46(11):1160–1165.
5. Rheinbay E, Parasuraman P, Grimsby J, et al. Recurrent and functional regulatory mutations in breast cancer. *Nature*. 2017;547(7661):55–60.
6. Lawrence MS, Stojanov P, Mermel CH, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*. 2014;505(7484):495–501.
7. Rowland A, Dias MM, Wiese MD, et al. Meta-analysis comparing the efficacy of anti-EGFR monoclonal antibody therapy between KRAS G13D and other KRAS mutant metastatic colorectal cancer tumours. *Eur J Cancer*. 2016;55:122–130.
8. Katainen R, Dave K, Pitkanen E, et al. CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat Genet*. 2015;47(7):818–821.
9. Saunders CT, Wong WS, Swamy S, et al. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics*. 2012;28(14):1811–1817.
10. The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*. 2012;487(7407):330–337.
11. Wala JA, Bandopadhayay P, Greenwald NF, et al. SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res*. 2018;28(4):581–591.
12. Narzisi G, Corvelo A, Arora K, et al. Genome-wide somatic variant calling using localized colored de Bruijn graphs. *Commun Biol*. 2018;1(1):20.
13. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. *Nature*. 2013;500(7463):415–421.
14. Forbes SA, Beare D, Gunasekaran P, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res*. 2015;43(D1):D805–D811.
15. Giannakis M, Mu XJ, Shukla SA, et al. Genomic correlates of immune-cell infiltrates in colorectal carcinoma. *Cell Reports*. 2016;15(4):857–865.
16. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucl Acids Res*. 2014;42(D1):D980–D985.
17. Plazzer JP, Sijmons RH, Woods MO, et al. The InSiGHT database: utilizing 100 years of insights into Lynch syndrome. *Fam Cancer*. 2013;12(2):175–180.
18. Haradhvala NJ, Kim J, Maruvka YE, et al. Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair. *Nat Commun*. 2018;9(1):1746.
19. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60, 706 humans. *Nature*. 2016;536(7616):285–291.
20. Ali M, Kim H, Cleary S, et al. Characterization of mutant MUTYH proteins associated with familial colorectal cancer. *Gastroenterology*. 2008;135(2):499–507.
21. Pilati C, Shinde J, Alexandrov LB, et al. Mutational signature analysis identifies MUTYH deficiency in colorectal cancers and adrenocortical carcinomas. *J Pathol*. 2017;242(1):10–15.
22. Vinagre J, Almeida A, Populo H, et al. Frequency of TERT promoter mutations in human cancers. *Nat Commun*. 2013;4:2185.
23. Huang FW, Hodis E, Xu MJ, et al. Highly recurrent TERT promoter mutations in human melanoma. *Science*. 2013;339(6122):957–959.
24. Killela PJ, Reitman ZJ, Jiao Y, et al. TERT promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal. *Proc Natl Acad Sci U S A*. 2013;110(15):6021–6026.
25. Cruvinel-Carloni A, Yamane L, Scapulatempo-Neto C, et al. Absence of TERT promoter mutations in colorectal precursor lesions and cancer. *Genet Mol Biol*. 2018;41(1):82–84.
26. Mularoni L, Sabarinathan R, Deu-Pons J, et al. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol*. 2016;17(1):128.
27. Fu Y, Liu Z, Lou S, et al. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol*. 2014;15(10):480.
28. Wang J, Zhuang J, Iyer S, et al. Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. *Nucleic Acids Res*. 2013;41(D1):D171–D176.
29. Hyvarinen AK, Pohjoismaki JL, Holt IJ, et al. Overexpression of MTERFD1 or MTERFD3 impairs the completion of mitochondrial DNA replication. *Mol Biol Rep*. 2011;38(2):1321–1328.
30. Reznik E, Miller ML, Şenbabaoğlu Y, et al. Mitochondrial DNA copy number variation across human cancers. *eLife*. 2016;5:e10769.
31. Rahman S, Magnussen M, León TE, et al. Activation of the LMO2 oncogene through a somatically acquired neomorphic promoter in T-cell acute lymphoblastic leukemia. *Blood*. 2017;129(24):3221–3226.
32. Mansour MR, Abraham BJ, Anders L, et al. Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science*. 2014;346(6215):1373–1377.
33. Groschel S, Sanders MA, Hoogenboezem R, et al. A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell*. 2014;157(2):369–381.
34. Abraham BJ, Hnisz D, Weintraub AS, et al. Small genomic insertions form enhancers that misregulate oncogenes. *Nat Commun*. 2017;8:14385.
35. Harrow J, Frankish A, Gonzalez JM, et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res*. 2012;22(9):1760–74.
36. The Encode Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489(7414):57–74.
37. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics*. 2015;31(16):2745–2747.
38. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet*. 2013;76(1):7.20.1–7.20.41.
39. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res*. 2003;31(13):3812–3814.