A comprehensive analysis of cancer-driving mutations and genes in kidney cancer

CHENGMEI LONG[1*], JINBO JIAN[2*], XINCHANG LI[1], GONGXIAN WANG[3] and JINGEN WANG[4]

[1]Department of Organ Transplantation, Jiangxi Provincial People's Hospital, School of Medicine, Nanchang University, Nanchang, Jiangxi 330006; [2]Department of Oncology, Binzhou Medical University Hospital, Binzhou, Shandong 256603; [3]Department of Urology, The First Affiliated Hospital of Nanchang University; [4]Department of Urology, Jiangxi Provincial People's Hospital, School of Medicine, Nanchang University, Nanchang, Jiangxi 330006, P.R. China

Abstract. An accumulation of driver mutations is important for cancer formation and progression, and leads to the disruption of genes and signaling pathways. The identification of driver mutations and genes has been the subject of numerous previous studies. The present study was performed to identify cancer-driving mutations and genes in renal cell carcinoma (RCC), prioritizing noncoding variants with a high functional impact, in order to analyze the most informative features. Sorting Intolerant From Tolerant (SIFT), Polymorphism Phenotyping version 2 (Polyphen2) and MutationAssessor were applied to predict deleterious mutations in the coding genome. OncodriveFM and OncodriveCLUST were used to detect potential driver genes and signaling pathways. The functional impact of noncoding variants was evaluated using Combined Annotation Dependent Depletion, FunSeq2 and Genome-Wide Annotation of Variants. Noncoding features were analyzed with respect to their enrichment of high-scoring variants. A total of 1,327 coding mutations in clear cell RCC, 258 in chromophobe RCC and 1,186 in papillary RCC were predicted to be deleterious by all three of MutationAssessor, Polyphen2 and SIFT. In total, 77 genes were positively selected by OncodriveFM and 1 by OncodriveCLUST, 45 of which were recurrently mutated genes. In addition, 10 signaling pathways were recurrently mutated and had a high functional impact bias (FM bias), and 31 novel signaling pathways with high FM bias were identified. Furthermore, noncoding regulatory features and conserved regions contained numerous high-scoring variants, and expression, replication time, GC content and recombination rate were positively correlated with the densities of high-scoring variants. In conclusion, the present study identified a list of cancer-driving genes and signaling pathways, features like regulatory elements, conserved regions, replication time, expression, GC content and recombination rate are major factors that affect the distribution of high-scoring non-coding mutations in kidney cancer.

Introduction

Rapid advancements in sequencing technology and its wide applications have identified hundreds of thousands of mutations in cancer genomes; a small fraction of these mutations, termed drivers, are critical for carcinogenesis and are able to confer a growth advantage to tumor cells by affecting driver genes (1,2). The detection of these cancer-driving mutations and genes has been the focus of numerous cancer genomic studies (3-6). Various computational approaches have been developed to prioritize deleterious cancer mutations, including the Sorting Intolerant From Tolerant (SIFT) algorithm (7), Polymorphism Phenotyping version 2 (PolyPhen2) tool (8) and MutationAssessor (9). The majority of these programs rely on the assumption that coding mutations that affect functionally important residues, as inferred from evolutionary conservation and protein domain analysis, are more likely to be deleterious (10). With regard to the identification of cancer-driving genes, common approaches search for genes that are recurrently mutated relative to the background mutation rate in a cohort of cancer cases, and include MutSigCV (2) and MuSiC (11). Whereas few driver genes are recurrently mutated in cancer tissue samples, other cancer drivers are mutated in a small fraction (<1%) of tumors (12). Therefore, methods that are able to classify driver genes independently of mutation frequency are required. OncodriveFM detects genes with a bias towards the accumulation of variants with high functional impacts, as evaluated by SIFT, PolyPhen2 and MutationAssessor (13). Another program, OncodriveCLUST, identifies genes with a significant bias towards gain-of-function mutations that are clustered within the protein-coding sequence, based on the knowledge that the clustering of gain-of-function mutations in specific protein regions provides an adaptive advantage to cancer cells, and is consequently positively selected for during

Correspondence to: Dr Jingen Wang, Department of Urology, Jiangxi Provincial People's Hospital, School of Medicine, Nanchang University, 13th Floor, South Building, 152 Aiguo Road, Nanchang, Jiangxi 330006, P.R. China
E-mail: drwang@jx163.com

*Contributed equally

tumoral evolution (14). At present, 547 cancer-associated genes are annotated in the Catalogue of Somatic Mutations in Cancer (COSMIC) database (15).

Although cancer mutations in the coding genome have been well studied, the majority of those in noncoding regions, which comprise 98% of the genome, remain poorly understood due to lack of functional information (16). An increasing number of noncoding pathogenic variants have been detected and annotated, including a large number of disease- or trait-associated single nucleotide polymorphisms detected in genome-wide association studies, preferentially within enhancers, exons and mRNA promoters (17). Therefore, it is important to develop reliable and efficient computational tools for evaluating the functional effects of noncoding variants. The completion of high-throughput projects, including the Encyclopedia of DNA Elements (ENCODE), has provided genome-wide mapping of histone modification and DNase I hypersensitive sites, formaldehyde-assisted isolation of regulatory elements, transcription factor binding sites (TFBS), and RNA sequencing and replication timing data for several cell lines, which has enabled the functional annotation of variants in the noncoding portion of the human genome (18). Various studies have utilized these data to predict and score the functionalities of noncoding variants. For example, in a study by Kircher *et al* (19), the annotations of fixed or almost fixed derived alleles observed in humans were contrasted with those of simulated *de novo* variants, and the Combined Annotation Dependent Depletion (CADD) tool was developed using a support vector machine. The application of CADD was able to effectively differentiate 14.7 million high-frequency human-derived alleles from 14.7 million simulated variants. Fu *et al* (20) developed the computational framework FunSeq2, which has processed large-scale genomic data (including 1000 Genomes and ENCODE data) and cancer resources, and used a high-throughput variant prioritization pipeline to annotate and prioritize somatic mutations. Genome-Wide Annotation of Variants (GWAVA) uses the regulatory mutations annotated in the Human Gene Mutation Database (21) and a combination of regulatory features, genic context and genome-wide properties to construct three random forest classifiers to score noncoding variants (22).

The current study was conducted to analyze the somatic mutations detected by whole-genome sequencing of 14 clear-cell renal cell carcinoma (ccRCC) tissue samples and exome sequencing of 106 ccRCC tissue specimens (23), 65 paired chromophobe renal cell carcinoma (chRCC) tissue samples and 100 paired papillary renal cell carcinoma (PRCC) tissue samples. OncodriveFM and OncodriveCLUST were used to identify the driver genes and signaling pathways that exhibited positive selection in kidney cancer. In addition, the scoring systems CADD, FunSeq2 and GWAVA were implemented in order to functionally annotate somatic variants in the noncoding genome. The enrichment of high-scoring variants for a wide range of noncoding features was also examined to identify the features that most contribute to the functionalities of noncoding cancer mutations.

Materials and methods

Cancer mutation data. Somatic mutations that had been detected using whole-genome sequencing of 14 pairs of ccRCC and normal tissue specimens and exome sequencing of 106 paired ccRCC tissue samples were obtained from the supplementary data of the study by Sato *et al* (23). Data for chRCC and PRCC mutations that had been identified using exome-sequencing of 65 and 100 paired chRCC and PRCC tissue samples by Lawrence *et al* (2) were also obtained.

Prediction of the functional effects of somatic mutations on cancer genes and signaling pathways. The functional impacts of somatic mutations in the coding genome were predicted using SIFT, Polyphen2 and MutationAssessor version 3. Variants were considered deleterious based on the following criteria: SIFT score <0.05; the presence of non-benign variants in the HumDiv and HumVar predictions of Polyphen2; and a MutationAssessor score >1.9. Cancer genes and signaling pathways were predicted using OncodriveFM 0.0.1 and OncodriveCLUST 0.4.1 (http://www.intogen.org/analysis/mutations), respectively, with all parameters set to default. Genes and signaling pathways in which P<0.05 were regarded as cancer gene and signaling pathway candidates. Gene Ontology (GO) enrichment analysis was performed for all the driver candidates (http://geneontology.org/) (24).

Genome-wide data resources. Human genome annotation data were obtained from Gencode V21, including protein coding genes, exons, introns, untranslated regions (UTRs) and noncoding exons (25). Evolutionarily conserved bases were identified using the recently published analysis of 46 mammalian genomes (26). The evolutionarily conserved secondary RNA structures were obtained from the study by Smith *et al* (27), in which they were predicted using comparative structure algorithms based on numerous genomes. Promoters, which are defined as regions 2.5 Kb from transcription start sites, were generated by the Gerstein lab (28). Histone acetylation and methylation data of cluster of differentiation (CD)4+ T cell lines were acquired from Wang *et al* (29) and Barski *et al* (30), respectively, and all coordinates from human genome (hg)18 assembly were converted into hg19 using the University of California, Santa Cruz (UCSC) LiftOver program (26). Conserved TFBS from in the human/mouse/rat alignment were obtained from UCSC directly (26). A wavelet-smoothed, weighted average signal with high and low values that indicate early and late replication during the S phase, respectively (http://genome.ucsc.edu/; ENCODE, Repli-seq track), was utilized (18). Genome-wide replication timing of the HepG2 cell line (26), which was used as there is no replication time data for kidney cancer cell line was mapped to protein coding genes and long noncoding RNAs (lncRNA), and an early-to-late ratio was calculated as $(G_1b+S_1)/(S_4+G_2)$, in which G_1b, S_1, S_4 and G_2 refer to replication timing values of the cell cycle fractions of G1b, S1, S4 and G2, respectively. If this ratio was >1, genes were considered to be replicated early, while late-replicated genes exhibited an early-to-late ratio <1. Recombination rates were obtained from The International HapMap Project (http://hapmap.ncbi.nlm. nih.gov/) and averaged over 1-Kb non-overlapping windows across the genome (31). Those windows with recombination rate >4.0 were considered high and those with a recombination rate <0.5 were considered low. The GC content refers to the fraction of G or C residues per 1-Kb window, and 1-Kb windows with a

GC fraction >50% or <30% were regarded as high or low GC regions, respectively (26).

RNA-seq data (GEO accession no. GSE55572) from 6 human embryonic kidney (HEK)293 T cells were obtained from the study by Schwartz *et al* (32) and reads were aligned to the hg19 genome using TopHat2 version 2.0.13 (33). Read counts were calculated using bedtools version 2.22.1 for each lncRNA and protein-coding gene (34). Expression levels were calculated by counting the number of reads per Kb per million reads (RPKM) and averaged from the 6 HEK293T cell lines for each protein coding gene and lncRNA. Genes with a RPKM >20 or <0.25 were defined as having high and low levels of gene expression, respectively.

Cancer lncRNAs containing 25 lncRNAs are a collection of mammalian long noncoding RNAs that have been experimentally demonstrated to be associated with a variety of cancer types. A list of cancer census genes was obtained from COSMIC version 71 (15).

Analysis of noncoding variants. In total, 70,659 noncoding variants detected by whole-genome sequencing of 14 paired ccRCC tissue samples were scored for deleteriousness using CADD v1.0 (http://cadd.gs.washington.edu), FunSeq2.1.2 (http://funseq2.gersteinlab.org) and GWAVA v1.0 (https://www.sanger.ac.uk/sanger/StatGen_Gwava), and all parameters were set to default. A total of 10,000 high-scoring noncoding variants predicted by CADD, FunSeq2 and GWAVA were intersected, and 1,454 variants (14.54%) that were scored as high by all three approaches were obtained. Next, 1,454 high-scoring variants identified by CADD, FunSeq2, GWAVA individually and in combination were selected and mapped onto various noncoding features. The density of high-scoring noncoding variants was calculated as variants/Mb for each feature.

Correlation analyses were performed to examine the associations between the densities of high-scoring noncoding variants and expression levels, replication time, GC content and recombination rate. Genes and lncRNAs with expression and replication times, as well as the 1-Kb windows with GC content and average recombination rates, were sorted and divided into non-overlapping 100 Mb intervals based on expression levels, replication times, GC content and average recombination rate respectively. The density of high-scoring noncoding variants was computed for each interval, the correlations between the average expression levels, replication time, GC content and recombination rate, and the densities of high-scoring noncoding variants were evaluated using Pearson correlation in R 3.2.0.

Statistical analysis. Data were presented as the mean. Variation between groups was examined using the Fisher's exact test. Correlation analysis was conducted with Pearson correlation in R 3.2.0 and P<0.05 was considered to indicate a statistically significant difference.

Results

Catalogue of somatic mutations. In total, 76,595 somatic mutations were obtained from Sato *et al* (23), comprising 71,424 mutations generated by whole-genome sequencing of 14 ccRCC tissue samples and 5,171 mutations detected by whole-exome sequencing of 106 ccRCC specimens. Among these, 72,871 were single-nucleotide variants (SNVs) and 3,724 were small insertions or deletions (indels). A total of 1,381 mutations detected by exome sequencing of 65 paired chRCC tissue samples included 1,287 SNVs and 94 indels; whereas 6,349 mutations were detected by exome sequencing of the 100 paired PRCC specimens, consisting of 5,489 SNVs, 677 indels and 180 dinitropyrenes. The fraction of mutations that were predicted to be deleterious varied greatly across cancer subtypes and prediction tools. The percentages of predicted deleterious mutations were as follows: 44.52% (2673/6004; SIFT), 50.45% (2393/4743; Polyphen2) and 39.18% (1941/4954; MutationAssessor) in ccRCC; 41.04% (2464/6330; SIFT), 55.91% (2110/3774; Polyphen2) and 44.47% (390/877; MutationAssessor) in chRCC; and 37.58% (519/1381; SIFT), 57.26% (469/819; Polyphen2) and 44.33% (1752/3952; MutationAssessor) in PRCC. In total, 1,327, 258 and 1,186 common SNVs were predicted to be deleterious by all three tools in ccRCC, chRCC and PRCC, respectively (Fig. 1A-C), and 838, 45 and 414 small indels introduced translational frameshifts, respectively, as predicted by SIFT. T>C/A>G, C>T/G>A and C>A/G>T accounted for 24.96, 23.16 and 17.86% of the variants in ccRCC, 9.34, 57.13 and 12.53% of the variants in chRCC and 17.31, 28.29 and 12.24% of the variants in PRCC, respectively. T>C/A>G, C>T/G>A and C>A/G>T were therefore the three most common transitions identified in kidney cancer tissues (Fig. 1D).

Cancer driver genes in kidney cancer. Predicted deleterious mutations were mapped onto established cancer genes that had been annotated in the COSMIC database, revealing that the coding regions of cancer genes had a significantly higher enrichment of deleterious mutations, in comparison with those of non-cancer genes in ccRCC (259.68 vs. 95.06 variants/Mb; $P<2.2x10^{-16}$), chRCC (46.23 vs. 18.93 variants/Mb; $P=4.15x10^{-6}$) and PRCC (195.78 vs. 87.69 variants/Mb; $P<2.2x10^{-16}$; Fisher's exact test; Fig. 1E). OncodriveFM and OncodriveCLUST were applied to identify the driver genes in ccRCC, chRCC and PRCC. In total, 44 genes were determined as driver candidates by OncodriveFM and 1 by OncodriveCLUST in ccRCC, 5 by OncodriveFM in chRCC and 33 by OncodriveFM in PRCC.

Cancer genes are usually subtype-specific in kidney cancer, and only 4 candidates, SET domain-containing 2 (SETD2), BRCA1-associated protein 1 (BAP1), GRB10-interacting GYF protein 2 (GIGYF2) and ubiquitin protein ligase E3 component N-recognin 4 (UBR4), exhibited overlap between ccRCC and PRCC. These results were compared with the 777 recurrently mutated genes identified by Sato *et al* using the ccRCC data, revealing that 45 cancer gene candidates were recurrently mutated. Certain genes among these are established cancer genes in ccRCC, including polybromo 1 (PBRM1), Von Hippel-Lindau tumor suppressor (VHL), SETD2, BAP1 and lysine demethylase 5C (KDM5C); however, 4 cancer genes originally established in other cancer types were identified as drivers in ccRCC, including the following: AT-rich interaction domain 1A (ARID1A) of clear cell ovarian carcinoma; lysine methyltransferase 2C (MLL3) of medulloblastoma; BCR, RhoGEF and GTPase-activating protein (BCR) of chronic myeloid leukemia, acute lymphoblastic leukemia and acute myeloid leukemia; and A-Kinase anchoring protein 9 (AKAP9)
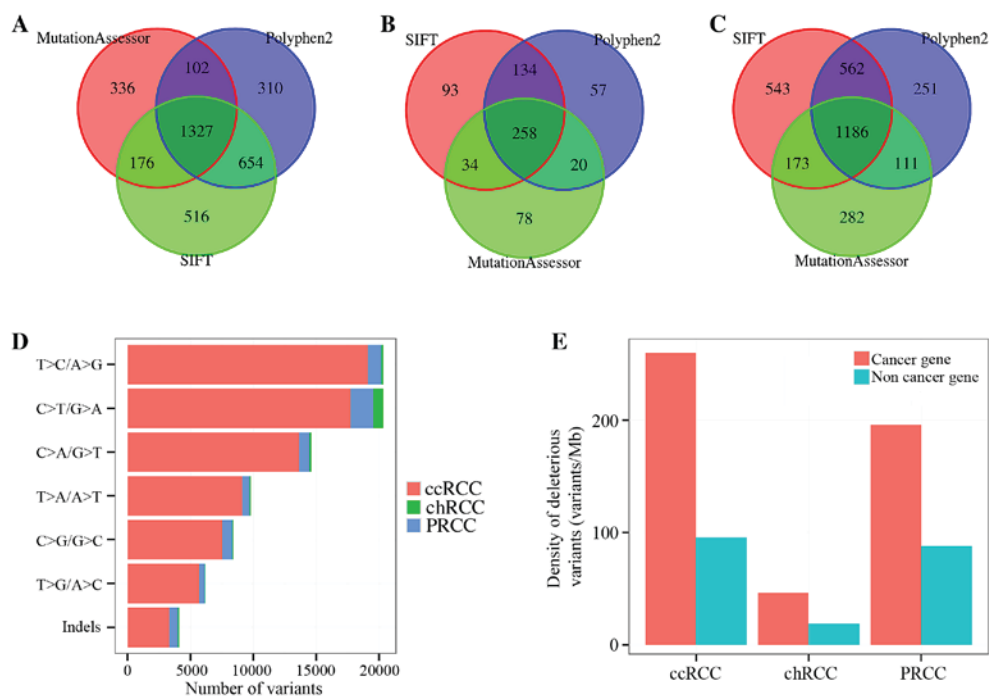
Figure 1. Venn diagram revealing the number of variants with deleterious effects predicted by SIFT, MutationAssesor and Polyphen2, and (A) the overlap between variants in clear cell renal cell carcinoma, (B) the overlap between variants in chromophobe renal cell carcinoma and (C) the overlap between variants in papillary renal cell carcinoma. (D) Mutation signatures in kidney cancer. (E) Densities of deleterious mutations in the coding regions of cancer genes and non-cancer genes. SIFT, Sorting Intolerant From Tolerant; polyphen2, Polymorphism Phenotyping version 2.

of papillary thyroid carcinoma (15). In addition, an important gene, transcription elongation factor B subunit 1 (TCEB1), was positively selected by all three algorithms. A total of 32 non-recurrently mutated genes were identified by Onco-driveFM, including adaptor-related protein complex 5 mu 1 subunit (AP5M1), chromodomain helicase DNA-binding protein 1 (CHD1), myosin heavy chain 11 (MYH11) and shugoshin-like 2 (SGOL2). CHD1 and MYH11 have been demonstrated to be involved in various cancer types (35-40).

Enrichment analysis of GO terms was performed for the 77 recurrently mutated cancer gene candidates, and 59 GO terms were determined with significant statistical evidence of P<0.05, including the regulation of metabolic processes (51 genes), positive regulation of metabolic processes (35 genes), positive regulation of macromolecular metabolic processes (29 genes), regulation of macromolecular metabolic processes (43 genes), positive regulation of cellular metabolic processes (29 genes), regulation of cellular processes (60 genes), chromosome organization (17 genes), regulation of biological processes (61 genes), negative regulation of biological processes (37 genes), regulation of cellular metabolic processes (43 genes), negative regulation of cellular processes (35 genes), positive regulation of biological processes (40 genes), chromatin modification (13 genes) and chromatin organization (14 genes).

Cancer-driving pathways in ccRCC. OncodriveFM analysis revealed 41 pathways with high FM bias in kidney cancer (Table I), including the following: Oxidative phosphorylation, spliceosome, RNA degradation, phagosome, legionellosis, HIF-1 signaling, leukocyte transendothelial migration, renal cell carcinoma, Wnt signaling and MAPK signaling pathways;

ubiquitin-mediated proteolysis pathway (UMPP); and pathways in cancer. Cancer pathways varied between kidney cancer subtypes, with 10 cancer pathways in ccRCC, 28 in chRCC and 5 in PRCC; only the Wnt signaling pathway and pathways in cancer exhibited overlap between PRCC and chRCC, and ccRCC and chRCC, respectively. The enriched mutational pathways obtained from Sato *et al* were compared with the results of the present study, revealing that 10 signaling pathways were positively selected by the two studies, including the following: UMPP, pathways in cancer, HIF-1 signaling pathway and the renal cell carcinoma pathway in ccRCC; and small cell lung cancer, P53 signaling pathway, mTOR pathway, prostate cancer, melanoma and PI3K-AKT pathway in chRCC. The remaining 31 pathways were novel signaling pathways with a high FM bias in kidney cancer.

Characterization of high-scoring variants and influential features in the noncoding genome. Previous studies have identified and annotated a number of noncoding pathogenic mutations (17,28). The present study therefore utilized three approaches (CADD, FunSeq2 and GWAVA) to functionally annotate noncoding variants that were detected by whole genome sequencing of 14 paired ccRCC tissue samples. Fig. 2A presents the density plots of the scores of all the noncoding variants predicted by CADD, FunSeq2 and GWAVA. The distribution of the scores differed between the three scoring systems. A total of 10,000 high-scoring noncoding variants were examined for intersections among them, and 1,454 (14.54%) variants were scored as high using all three approaches (Fig. 2B). Subsequently, the 1,454 high-scoring variants were selected from CADD, FunSeq2 and GWAVA analysis along with 1,454 common variants, in order to analyze

Table I. Cancer-driving signaling pathways as detected by OncodriveFM in kidney cancer.

A, Clear cell renal cell carcinoma

Pathway name	Pathway Identification number	Gene number	FM_Z score	P-value	Q-value
Oxidative phosphorylation	hsa00190	121	3.82	6.80×10^{-5}	1.90×10^{-3}
Spliceosome	hsa03040	125	3.54	1.97×10^{-4}	3.44×10^{-3}
RNA degradation	hsa03018	70	4.14	1.74×10^{-5}	7.98×10^{-4}
Ubiquitin-mediated proteolysis	hsa04120	137	6.64	1.60×10^{-11}	2.24×10^{-9}
Phagosome	hsa04145	147	3.08	1.05×10^{-3}	1.63×10^{-2}
Legionellosis	hsa05134	53	2.71	3.36×10^{-3}	4.71×10^{-2}
Pathways in cancer	hsa05200	326	4.08	2.28×10^{-5}	7.98×10^{-4}
HIF-1 signaling	hsa04066	110	3.59	1.67×10^{-4}	3.44×10^{-3}
Leukocyte transendothelial migration	hsa04670	114	3.55	1.94×10^{-4}	3.44×10^{-3}
Renal cell carcinoma	hsa05211	70	5.57	1.24×10^{-8}	8.69×10^{-7}

B, Papillary renal cell carcinoma

Pathway name	Pathway Identification number	Gene number	FM_Z score	P-value	Q-value
Wnt signaling pathway	hsa04310	152	2.88	1.97×10^{-3}	$8.94 \times 10^{-5}2$
Metabolic pathways	hsa01100	1160	2.81	2.46×10^{-3}	0.09
Pyrimidine metabolism	hsa00240	101	1.69	4.56×10^{-2}	0.98
Citrate cycle	hsa00020	30	2.44	7.30×10^{-3}	0.20
Viral myocarditis	hsa05416	68	3.60	1.57×10^{-4}	1.71×10^{-2}

C, Chromophobe renal cell carcinoma

Pathway name	Pathway Identification number	Gene number	FM_Z score	P-value	Q-value
Neurotrophin signaling pathway	hsa04722	119	5.38	3.63×10^{-8}	4.23×10^{-8}
Herpes simplex infection	hsa05168	182	6.96	1.71×10^{-12}	4.27×10^{-12}
Epstein-Barr virus infection	hsa05169	199	7.18	3.58×10^{-13}	1.25×10^{-12}
HTLV-I infection	hsa05166	260	6.01	9.02×10^{-10}	1.17×10^{-9}
Hepatitis C	hsa05160	131	6.71	9.80×10^{-12}	1.91×10^{-11}
Hepatitis B	hsa05161	147	7.39	7.34×10^{-14}	3.67×10^{-13}
Measles	hsa05162	134	6.83	4.18×10^{-12}	9.15×10^{-12}
Wnt signaling pathway	hsa04310	152	6.02	8.93×10^{-10}	1.17×10^{-9}
MAPK signaling pathway	hsa04010	257	7.41	6.15×10^{-14}	3.59×10^{-13}
Chronic myeloid leukemia	hsa05220	73	7.57	1.94×10^{-14}	1.70×10^{-13}
Non-small cell lung cancer	hsa05223	54	7.09	6.48×10^{-13}	2.06×10^{-12}
Small cell lung cancer	hsa05222	85	7.23	2.49×10^{-13}	9.68×10^{-13}
Focal adhesion	hsa04510	204	2.96	1.56×10^{-3}	1.71×10^{-3}
Cell cycle	hsa04110	124	6.96	1.71×10^{-12}	4.27×10^{-12}
Apoptosis	hsa04210	87	6.57	2.54×10^{-11}	4.67×10^{-11}
p53 signaling pathway	hsa04115	68	6.35	1.08×10^{-10}	1.79×10^{-10}
Transcriptional misregulation in cancer	hsa05202	180	5.78	3.69×10^{-9}	4.46×10^{-9}
Viral carcinogenesis	hsa05203	203	6.92	2.32×10^{-12}	5.41×10^{-12}
Pathways in cancer	hsa05200	326	8.17	1.53×10^{-16}	5.34×10^{-15}
Amyotrophic lateral sclerosis	hsa05014	53	6.24	2.22×10^{-10}	3.53×10^{-10}
Bladder cancer	hsa05219	42	6.15	3.84×10^{-10}	5.37×10^{-10}
mTOR signaling pathway	hsa04150	64	3.92	4.47×10^{-5}	5.05×10^{-5}

Table I. Continued.

Pathway name	Pathway identification number	Gene number	FM_Z score	P-value	Q-value
Huntington's disease	hsa05016	180	7.03	1.01×10^{-12}	2.94×10^{-12}
Thyroid cancer	hsa05216	29	6.17	3.37×10^{-10}	4.92×10^{-10}
Prostate cancer	hsa05215	88	7.98	7.14×10^{-16}	8.33×10^{-15}
Melanoma	hsa05218	71	7.46	4.40×10^{-14}	3.08×10^{-13}
Basal cell carcinoma	hsa05217	55	6.17	3.37×10^{-10}	4.92×10^{-10}
PI3K-Akt signaling pathway	hsa04151	338	5.90	1.81×10^{-9}	2.26×10^{-9}
Pancreatic cancer	hsa05212	70	7.35	9.55×10^{-14}	4.18×10^{-13}
Endometrial cancer	hsa05213	52	6.81	4.92×10^{-12}	1.01×10^{-11}
Glioma	hsa05214	65	8.07	3.60×10^{-16}	6.30×10^{-15}
Colorectal cancer	hsa05210	62	6.56	2.69×10^{-11}	4.70×10^{-11}

HIF-1, hypoxia-inducible factor 1; HTLV-I, human T-lymphotropic virus I; MAPK, mitogen-activated protein kinase; p53, tumor protein 53; mTOR, mechanistic target of rapamycin; PI3K-Akt, phosphoinositide 3-kinase-protein kinase B.
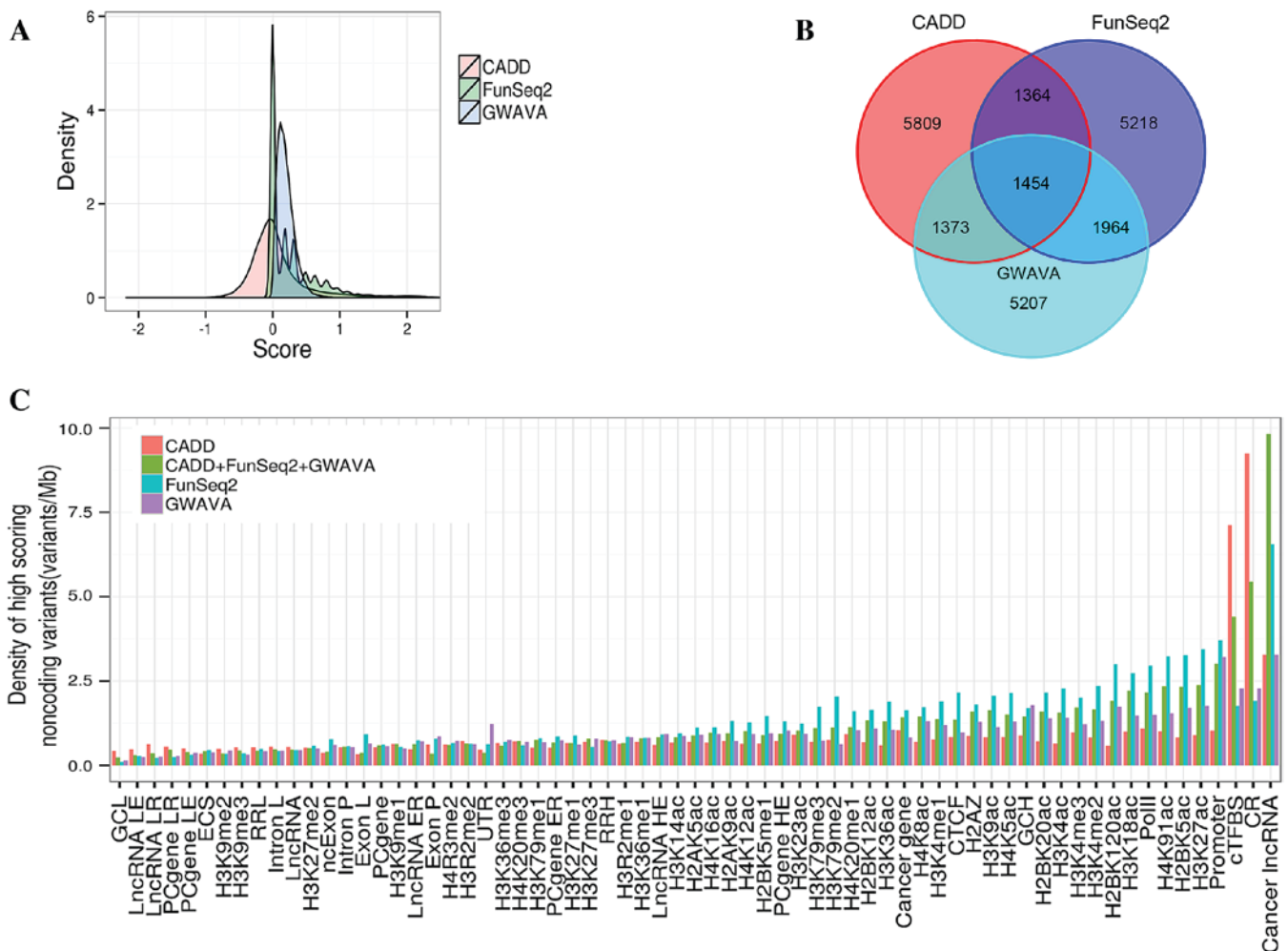

Figure 2. (A) Density plots of the scores of all noncoding variants, as predicted by CADD, FunSeq2 and GWAVA. (B) The highest 10,000 scoring noncoding variants, as predicted by each method, and the overlap between them. (C) Barplot presenting the densities of the 1,454 overlapping high-scoring noncoding variants, as predicted by each of the methods individually as well as in combination, in various noncoding features. CADD, Combined Annotation Dependent Depletion; GWAVA, Genome-Wide Annotation of Variants; GCL, GC content low in 1-Kb windows; lncRNA, long noncoding RNA; LE, low expression levels; LR, late replicated; PCgene, protein-coding gene; RRL, replication rate low in 1-Kb windows; Intron L, introns of lncRNAs; ncExon, non coding exon; Intron P, introns of PCgenes; Exon L, exons of lncRNAs; Exon P, exons of PCgenes; ER, early replicated; UTR, untranslated region; RRH, replication rate high in 1-Kb windows; HE, high expression levels; GCH, GC content high in 1-Kb windows; cTFBS, conserved transcription factor binding sites; CR, conserved region.
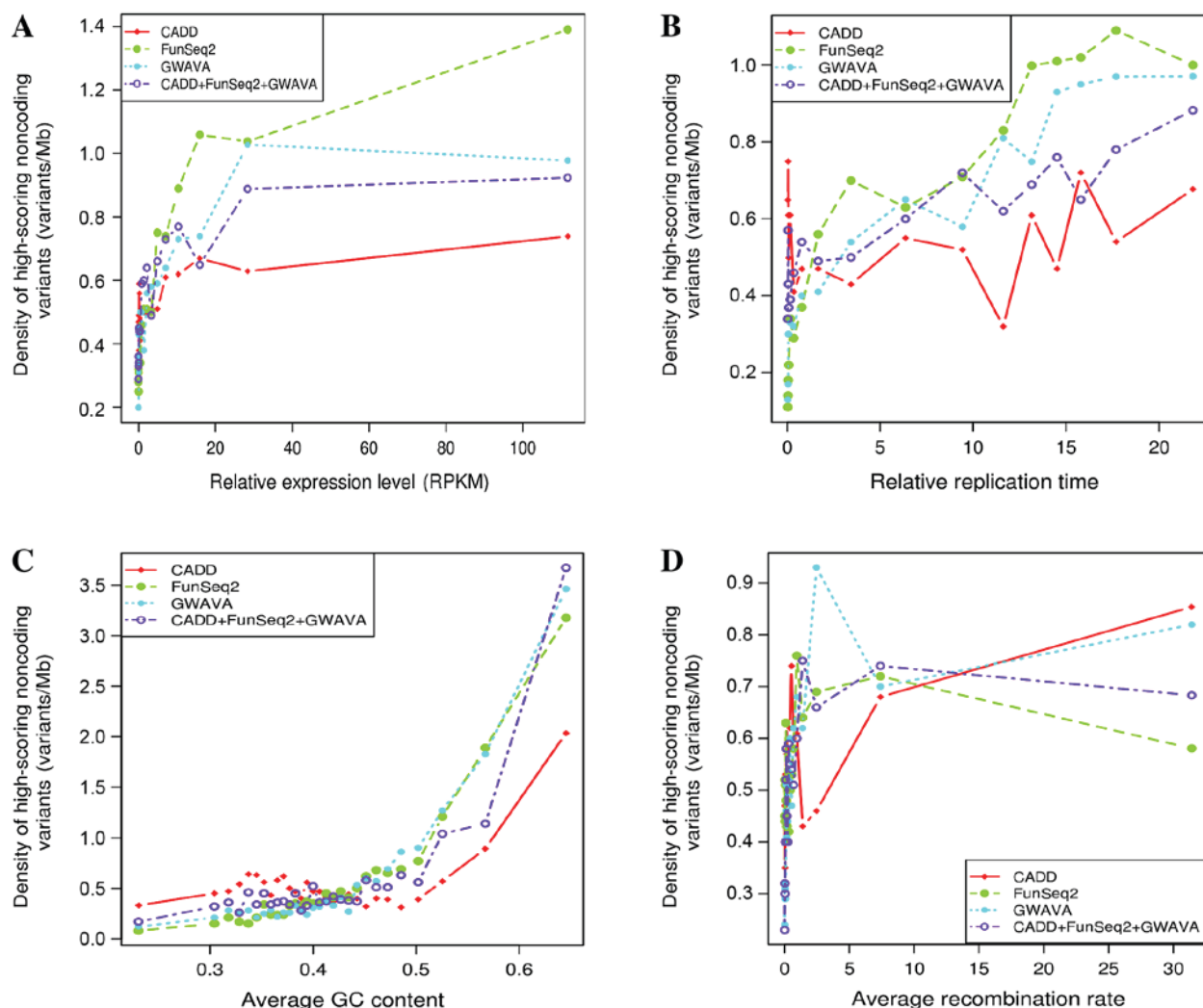
Figure 3. (A) Correlation between gene expression levels (RPKM) and the densities of high-scoring variants. (B) Correlation between replication time calculated as $(G_1b+S_1)/(S_4+G_2)$ and the densities of high-scoring variants. (C) Correlation between GC content (representing the fraction of GC bases in 1-Kb windows) and the densities of high-scoring variants. (D) Correlation between average recombination rate and the densities of high-scoring variants. CADD, Combined Annotation Dependent Depletion; GWAVA, Genome-Wide Annotation of Variants; RPKM, reads per Kb per million reads.

the distribution of these variants and evaluate the most important noncoding features for their formation. Fig. 2C shows that conserved regions and regulatory elements contained higher densities of high-scoring variants, as predicted by all scoring methods individually and in combination, including conserved regions, conserved TFBS, promoters, H3K27ac, H2BK5ac, H4K91ac, PolII, H3K18ac, H2BK120ac, H3K4me2 and H3K4me3. By contrast, repressive histone modifications, including H3K9me3 and H3K9me2, evolutionarily conserved structures, noncoding exons and UTRs ranked low with respect to the enrichment of high-scoring noncoding variants. In general, cancer genes contained a 1- to 3-fold significant enrichment of high-scoring regions in comparison with protein coding genes (common variants, $P=1.669 \times 10^{-9}$; CADD, $P=4.055 \times 10^{-5}$; FunSeq2, $P=2.148 \times 10^{-12}$; GWAVA, $P=0.033$; Fisher's exact test).

The present study also identified that features including expression levels, replication time, GC content and recombination rate are important for the densities of high-scoring mutations in the noncoding genome. For instance, highly expressed protein coding genes and lncRNAs are significantly

more enriched with high-scoring variants, compared with those that are expressed at low levels for common variants (protein coding genes, $P=4.293 \times 10^{-13}$; lncRNA, $P=4.414 \times 20^{-6}$), CADD (protein coding genes, $P=10 \times 3.777^{-3}$; lncRNA, $P=0.3501$), Funseq2 (protein coding genes, $P=2.2 \times 10^{-16}$; lncRNA, $P=6.576 \times 10^{-8}$) and GWAVA (protein coding genes, 2.2×10^{-16}; lncRNA, $P=2.302 \times 10^{-9}$; Fisher's exact test). Early-replicated protein coding genes and lncRNAs were significantly more enriched with high-scoring variants relative to late-replicated protein coding genes and lncRNAs, for common variants (protein coding genes, $P=5.84 \times 10^{-6}$; lncRNA $P=4.684 \times 10^{-4}$), CADD (protein coding genes, $P=0.3822$; lncRNA, $P=0.07757$), FunSeq2 (protein coding genes, $P=2.2 \times 10^{-16}$; lncRNA, $P=1.602 \times 10^{-12}$) and GWAVA (protein coding genes, $P=2.2 \times 10^{-16}$; lncRNA, $P=3.171 \times 10^{-9}$; Fisher's exact test). GC-rich regions that contained a high fraction of GC content possessed a significantly higher density of high-scoring variants as compared with low GC regions for common variants ($P<2.2 \times 10^{-16}$), CADD ($P=1.328 \times 10^{-6}$), FunSeq2 ($P=2.2 \times 10^{-16}$) and GWAVA ($P=2.2 \times 10^{-16}$; Fisher's exact test). Regions with a high average recombination rate

possessed a higher density of high-scoring variants compared with regions with a low average recombination rate for common variants (P=1.357x10^{-5}), CADD (P=2.583x10^{-3}), FunSeq2 (P=2.063x10^{-3}) and GWAVA (P=3.834x10^{-6}; Fisher's exact test). In addition, the expression levels, replication time, GC content and recombination rate exhibited positive correlations with the densities of high-scoring variants (Fig. 3A-D). For example, there was a positive correlation between gene expression levels (RPKM) and the density of high-scoring variants for CADD (r=0.69; P=0.0021), FunSeq2 (r=0.78; P=2x10^{-4}), GWAVA (r=0.69; P=0.0022) and the three methods combined (r=0.65; P=0.0048; Fig. 3A). Replication time was also correlated with the densities of high-scoring noncoding variants, as predicted by CADD (r=0.12; P=0.6539), FunSeq2 (r=0.93, P=8.32x10^{-8}), GWAVA (r=0.95; P=3.369x10^{-9}) and the three methods combined (r=0.91; P=3.519x10^{-7}; Fig. 3B). All these findings suggest that conserved regions, regulatory elements, high expression levels, early replication time, high GC content and high recombination rate are important features that affect the functionalities of noncoding variants in kidney cancer.

Discussion

The current study performed a full analysis of the somatic mutations generated by whole-genome and -exome sequencing of kidney cancer samples, revealing 1,327, 258 and 1,186 deleterious coding variants in ccRCC, chRCC and PRCC, respectively, predicted by SIFT, Polyphen2 and MutationAssessor. Implementation of OncodriveFM and OncodriveCluster enabled the identification of 77 cancer gene candidates and 41 cancer signaling pathways. Among them are established kidney cancer genes, including PBRM1, VHL, SETD2, BAP1 and KDM5C (41,42). The majority of candidates (45/77, 58.44%) were recurrently mutated genes; however, 32 driver gene candidates were not frequently mutated in kidney cancer tissue samples. An important gene, transcription elongation factor B subunit 1 (TCEB1), was positively selected by all three algorithms. TCEB1 encodes elongin C, which is a subunit of the heterotrimeric RNA polymerase II elongation factor complex that potently induces mRNA elongation (43). TCEB1 is overexpressed and amplified in prostate cancer, enhancing the cellular growth rate, whereas TCEB1 silencing decreases the invasion and growth of prostate cancer cells (44). The oncogenic role of TCEBI has been reported in ccRCC tumors containing TCEB1 mutations, which exhibited increased expression levels of hypoxia-inducible factor (HIF)-1α, a gene that is implicated to be dysregulated in various cancer-associated processes, including vascularization, angiogenesis, energy metabolism, cell survival and tumor invasion (23). The present study also identified RB transcriptional corepressor 1 (RB1), mechanistic target of rapamycin (MTOR), phosphatase and tensin homolog (PTEN) and tumor protein P53 (TP53) as specifically predicted cancer genes in chRCC. RB1, TP53 and PTEN are established cancer genes and may drive the formation and development of chRCC (45-47). In addition, GO term analysis of these 77 genes revealed that they are enriched in GO terms, including the regulation of metabolic process, regulation of cellular processes, chromosome organization, regulation of biological processes, chromatin modification and organization, all of which are involved in the pathogenesis of ccRCC (48-50).

In addition, 41 cancer-associated signaling pathways with high FM bias were identified, 10 of which are significant mutational pathways, including UMPP, pathways in cancer, and HIF-1 signaling and renal cell carcinoma pathways. Alterations in UMPP are associated with the overexpression of HIF-1α and HIF-2α, which are two crucial hypoxia regulatory factors in the HIF-1 signaling pathway; therefore, alterations in UMPP may contribute to the pathogenesis of ccRCC via the activation of the HIF-1 signaling pathway, which is important role in ccRCC tumorigenesis (42,51). Cancer signaling pathways were also positively selected by OncodriveFM in the present study, including pathways in cancer and renal cell carcinoma pathway, primarily consisting of the VHL, TCEB1, TCEB2 and HIF-1α pathways (52). In addition, 31 novel signaling pathways were identified to be associated with kidney cancer in the current study. An advantage of OncodriveFM and OncodriveCluster is that these two tools identify those genes and signaling pathways that accumulate variants with a high functional impact independently of cancer mutation frequency, enabling the identification of potential cancer genes and signaling pathways that are not highly mutated in cancer (13,14).

Application of the CADD, FunSeq2 and GWAVA scoring systems enables the quantitative evaluation of the functional effects of noncoding variants, and further analysis of the features that are important for their formation. The present study revealed that conserved regions, promoters and TFBS as well as numerous histone modifications were enriched with high-scoring variants, and that certain histone modifications were hallmarks of regulatory elements that are enriched with high-scoring variants. These histone markers included H3K4ac, H3K9ac, H2BK20ac, H2BK120ac, H3K18ac, H4K91ac and H2BK5ac, which are associated with transcription start sites as well as H3K27ac, H3K4me2 and H3K4me3, which mark active enhancers or promoters (29). By contrast, repressive histone markers, including H3K9me2 and H3K9me3, were characterized as markers of transcriptional repression, which accounts for their low enrichment of high-scoring variants (29).

In addition, the present study demonstrated that the density of high-scoring noncoding variants was strongly correlated with expression levels, replication time, GC content and recombination rate. Replication timing is an important epigenetic factor, and refers to the order in which segments of DNA along the length of a chromosome are duplicated (53). Regions of constant early replication are associated with high gene density and gene expression levels, GC content and cytosine-phosphate-guanine density, as well as vertebrate non-exonic conservation (54). GC-rich regions contain a high density of oncogenes or tumor suppressor genes (55). Recombination between homologous DNA sequences may lead to rearrangements, including a loss of heterozygosity, deletions, duplications, inversions and gene fusion; therefore, a higher recombination rate may be an important cause of the development of harmful mutations in the noncoding genome and predisposition to cancer (56). This supports the hypothesis that high expression levels, early replication time, high GC content and high recombination rate characterize cancer-implicated

regions within the noncoding genome, in which variants are more likely to be pathogenic.

In conclusion, the present study identified a set of cancer-associated genes and signaling pathways in data from kidney cancer tissue samples. Features including conservation, regulatory elements, replication time, expression levels, GC content and recombination rate may be important for the functional impact of noncoding mutations in renal cell carcinoma.

References

1. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz L Jr and Kinzler KW: Cancer genome landscapes. Science 339: 1546-1558, 2013.
2. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, et al: Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499: 214-218, 2013.
3. Davis CF, Ricketts CJ, Wang M, Yang L, Cherniack AD, Shen H, Buhay C, Kang H, Kim SC, Fahey CC, et al: The somatic genomic landscape of chromophobe renal cell carcinoma. Cancer Cell 26: 319-330, 2014.
4. Cancer Genome Atlas Research Network: Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature 499: 43-49, 2013.
5. Cancer Genome Atlas Network: Comprehensive molecular characterization of human colon and rectal cancer. Nature 487: 330-337, 2012.
6. Cancer Genome Atlas Network: Comprehensive molecular portraits of human breast tumours. Nature 490: 61-70, 2012.
7. Sim NL, Kumar P, Hu J, Henikoff S, Schneider G and Ng PC: SIFT web server: Predicting effects of amino acid substitutions on proteins. Nucleic Acids Res 40: W452-W457, 2012.
8. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS and Sunyaev SR: A method and server for predicting damaging missense mutations. Nat Methods 7: 248-249, 2010.
9. Reva B, Antipin Y and Sander C: Predicting the functional impact of protein mutations: Application to cancer genomics. Nucleic Acids Res 39: e118, 2011.
10. Vitkup D, Sander C and Church GM: The amino-acid mutational spectrum of human genetic disease. Genome Biol 4: R72, 2003.
11. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, Mooney TB, Callaway MB, Dooling D, Mardis ER, et al: MuSiC: Identifying mutational significance in cancer genomes. Genome Res 22: 1589-1598, 2012.
12. Wood LD, Parsons DW, Jones S, Lin J, Sjöblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, et al: The genomic landscapes of human breast and colorecta l cancers. Science 318: 1108-1113, 2007.
13. Gonzalez-Perez A and Lopez-Bigas N: Functional impact bias reveals cancer drivers. Nucleic Acids Res 40: e169, 2012.
14. Tamborero D, Gonzalez-Perez A and Lopez-Bigas N: OncodriveCLUST: Exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics 29: 2238-2244, 2013.
15. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A, et al: COSMIC: Mining complete cancer genomes in the catalogue of somatic mutations in cancer. Nucleic Acids Res 39 (Database Issue): D945-D950, 2011.
16. Weinhold N, Jacobsen A, Schultz N, Sander C and Lee W: Genome-wide analysis of noncoding regulatory mutations in cancer. Nat Genet 46: 1160-1165, 2014.
17. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, et al: An atlas of active enhancers across human cell types and tissues. Nature 507: 455-461, 2014.
18. Rosenbloom KR, Sloan CA, Malladi VS, Dreszer TR, Learned K, Kirkup VM, Wong MC, Maddren M, Fang R, Heitner SG, et al: ENCODE data in the UCSC genome browser: Year 5 update. Nucleic Acids Res 41 (Database Issue): D56-D63, 2013.
19. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM and Shendure J: A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46: 310-315, 2014.
20. Fu Y, Liu Z, Lou S, Bedford J, Mu XJ, Yip KY, Khurana E and Gerstein M: FunSeq2: A framework for prioritizing noncoding regulatory variants in cancer. Genome Biol 15: 480, 2014.
21. Stenson PD, Mort M, Ball EV, Howells K, Phillips AD, Thomas NST and Cooper DN: Human gene mutation database: 2008 update. Genome Med 1: 13, 2009.
22. Ritchie GR, Dunham I, Zeggini E and Flicek P: Functional annotation of noncoding sequence variants. Nat Methods 11: 294-296, 2014.
23. Sato Y, Yoshizato T, Shiraishi Y, Maekawa S, Okuno Y, Kamura T, Shimamura T, Sato-Otsubo A, Nagae G, Suzuki H, et al: Integrated molecular analysis of clear-cell renal cell carcinoma. Nat Genetics 45: 860-867, 2013.
24. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al: Gene ontology: Tool for the unification of biology. The gene ontology consortium. Nat Genet 25: 25-29, 2000.
25. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al: GENCODE: The reference human genome annotation for the ENCODE project. Genome Res 22: 1760-1774, 2012.
26. Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, et al: The UCSC genome browser database: 2014 update. Nucleic Acids Res 42 (Database Issue): D764-D770, 2014.
27. Smith MA, Gesell T, Stadler PF and Mattick JS: Widespread purifying selection on RNA structure in mammals. Nucleic Acids Res 41: 8220-8236, 2013.
28. Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, Sboner A, Lochovsky L, Chen J, Harmanci A, et al: Integrative annotation of variants from 1092 humans: Application to cancer genomics. Science 342: 1235587, 2013.
29. Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh TY, Peng W, Zhang MQ and Zhao K: Combinatorial patterns of histone acetylations and methylations in the human genome. Nat Genet 40: 897-903, 2008.
30. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I and Zhao K: High-resolution profiling of histone methylations in the human genome. Cell 129: 823-837, 2007.
31. Pollard KS, Hubisz MJ, Rosenbloom KR and Siepel A: Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 20: 110-121, 2010.
32. Schwartz S, Mumbach MR, Jovanovic M, Wang T, Maciag K, Bushkin GG, Mertins P, Ter-Ovanesyan D, Habib N, Cacchiarelli D, et al: Perturbation of m6A writers reveals two distinct classes of mRNA methylation at internal and 5' sites. Cell Rep 8: 284-296, 2014.
33. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R and Salzberg SL: TopHat2 : Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14: R36, 2013.
34. Quinlan AR and Hall IM: BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841-842, 2010.
35. Burkhardt L, Fuchs S, Krohn A, Masser S, Mader M, Kluth M, Bachmann F, Huland H, Steuber T, Graefen M, et al: CHD1 is a 5q21 tumor suppressor required for ERG rearrangement in prostate cancer. Cancer Res 73: 2795-2805, 2013.
36. Huang S, Gulzar ZG, Salari K, Lapointe J, Brooks JD and Pollack JR: Recurrent deletion of CHD1 in prostate cancer with relevance to cell invasiveness. Oncogene 31: 4164-4170, 2012.
37. Alhopuro P, Karhu A, Winqvist R, Waltering K, Visakorpi T and Aaltonen LA: Somatic mutation analysis of MYH11 in breast and prostate cancer. BMC Cancer 8: 263, 2008.
38. Liu P, Tarlé SA, Hajra A, Claxton DF, Marlton P, Freedman M, Siciliano MJ and Collins FS: Fusion between transcription factor CBF beta/PEBP2 beta and a myosin heavy chain in acute myeloid leukemia. Science 261: 1041-1044, 1993.
39. Vickaryous N, Polanco-Echeverry G, Morrow S, Suraweera N, Thomas H, Tomlinson I and Silver A: Smooth-muscle myosin mutations in hereditary non-polyposis colorectal cancer syndrome. Br J Cancer 99: 1726-1728, 2008.
40. Alhopuro P, Phichith D, Tuupanen S, Sammalkorpi H, Nybondas M, Saharinen J, Robinson JP, Yang Z, Chen LQ, Orntoft T, et al: Unregulated smooth-muscle myosin in human intestinal neoplasia. Proc Natl Acad Sci USA 105: 5513-5518, 2008.

41. Varela I, Tarpey P, Raine K, Huang D, Ong CK, Stephens P, Davies H, Jones D, Lin ML, Teague J, *et al*: Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. Nature 469: 539-542, 2011.

42. Guo G, Gui Y, Gao S, Tang A, Hu X, Huang Y, Jia W, Li Z, He M, Sun L, *et al*: Frequent mutations of genes encoding ubiquitin-mediated proteolysis pathway components in clear cell renal cell carcinoma. Nat Genet 44: 17-19, 2011.

43. Aso T, Lane WS, Conaway JW and Conaway RC: Elongin (SIII): A multisubunit regulator of elongation by RNA polymerase II. Science 269: 1439-1443, 1995.

44. Jalava SE, Porkka KP, Rauhala HE, Isotalo J, Tammela TL and Visakorpi T: TCEB1 promotes invasion of prostate cancer cells. Int J Cancer 124: 95-102, 2009.

45. Xu XL, Singh HP, Wang L, Qi DL, Poulos BK, Abramson DH, Jhanwar SC and Cobrink D: Rb suppresses human cone-precursor-derived retinoblastoma tumours. Nature 514: 385-388, 2014.

46. Stracquadanio G, Wang X, Wallace MD, Grawenda AM, Zhang P, Hewitt J, Zeron-Medina J, Castro-Giner F, Tomlinson IP, Goding CR, *et al*: The importance of p53 pathway genetics in inherited and somatic cancer genomes. Nat Rev Cancer 16: 251-265, 2016.

47. Roa I, de Toro G, Fernández F, Game A, Muñoz S, de Aretxabala X and Javle M: Inactivation of tumor suppressor gene pten in early and advanced gallbladder cancer. Diagn Pathol 10: 148, 2015.

48. Buck MJ, Raaijmakers LM, Ramakrishnan S, Wang D, Valiyaparambil S, Liu S, Nowak NJ and Pili R: Alterations in chromatin accessibility and DNA methylation in clear cell renal cell carcinoma. Oncogene 33: 4961-4965, 2014.

49. Zaravinos A, Pieri M, Mourmouras N, Anastasiadou N, Zouvani I, Delakas D and Deltas C: Altered metabolic pathways in clear cell renal cell carcinoma: A meta-analysis and validation study focused on the deregulated genes and their associated networks. Oncoscience 1: 117-131, 2014.

50. Rankin EB, Fuh KC, Castellini L, Viswanathan K, Finger EC, Diep AN, LaGory EL, Kariolis MS, Chan A, Lindgren D, *et al*: Direct regulation of GAS6/AXL signaling by HIF promotes renal metastasis through SRC and MET. Proc Natl Acad Sci USA 111: 13373-13378, 2014.

51. Baldewijns MM, van Vlodrop IJ, Vermeulen PB, Soetekouw PM, van Engeland M and de Bruïne AP: VHL and HIF signalling in renal cell carcinogenesis. J Pathol 221: 125-138, 2010.

52. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H and Kanehisa M: KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 27: 29-34, 1999.

53. Rhind N and Gilbert DM: DNA replication timing. Cold Spring Harb Perspect Biol 5: a010132, 2013.

54. Hansen RS, Thomas S, Sandstrom R, Canfield TK, Thurman RE, Weaver M, Dorschner MO, Gartler SM and Stamatoyannopoulos JA: Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. Proc Natl Acad Sci USA 107: 139-144, 2010.

55. Watanabe Y, Abe T, Ikemura T and Maekawa M: Relationships between replication timing and GC content of cancer-related genes on human chromosomes 11q and 21q. Gene 433: 26-31, 2009.

56. Bishop AJ and Schiestl RH: Homologous recombination as a mechanism of carcinogenesis. Biochim Biophys Acta 1471: M109-M121, 2001.