Taylor & Francis
Taylor & Francis Group

RESEARCH PAPER

OPEN ACCESS   Check for updates

An integrated DNA and RNA variant detector identifies a highly conserved three base exon in the *MAP4K5* kinase locus

Małgorzata Kurkowiak [a,*], Giuseppa Grasso[b,^,*], Jakub Faktor [a,c], Lisa Scheiblecker[d], Małgorzata Winniczuk[a], Marcos Yebenes Mayordomo[a,b], J. Robert O'Neill[e], Bodil Oster[f], Borek Vojtesek[c], Ali Al-Saadi[b], Natalia Marek-Trzonkowska [a,g], and Ted R. Hupp [a,b]

aInternational Centre for Cancer Vaccine Science (ICCVS), University of Gdańsk, 80-822 Gdańsk, Poland; bUniversity of Edinburgh, Institute of Genetics and Molecular Medicine, Edinburgh Cancer Research Centre, Edinburgh, Scotland, UK; cResearch Centre for Applied Molecular Oncology, Masaryk Memorial Cancer Institute, Brno, Czech Republic; dInstitute of Pharmacology and Toxicology, University of Veterinary Medicine Vienna, 1210 Vienna, Austria; eCambridge Oesophagogastric Centre, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK; fQIAGEN Aarhus, Silkeborgvej 2, 8000 Aarhus, Denmark; gLaboratory of Immunoregulation and Cellular Therapies, Department of Family Medicine, Medical University of Gdańsk, Gdańsk, Poland

ABSTRACT
RNA variants that emerge from editing and alternative splicing form important regulatory stages in protein signalling. In this report, we apply an integrated DNA and RNA variant detection workbench to define the range of RNA variants that deviate from the reference genome in a human melanoma cell model. The RNA variants can be grouped into (i) classic ADAR-like or APOBEC-like RNA editing events and (ii) multiple-nucleotide variants (MNVs) including three and six base pair in-frame non-canonical unmapped exons. We focus on validating representative genes of these classes. First, clustered non-synonymous RNA edits (A-I) in the *CDK13* gene were validated by Sanger sequencing to confirm the integrity of the RNA variant detection workbench. Second, a highly conserved RNA variant in the *MAP4K5* gene was detected that results most likely from the splicing of a non-canonical three-base exon. The two RNA variants produced from the *MAP4K5* locus deviate from the genomic reference sequence and produce V569E or V569del isoform variants. Low doses of splicing inhibitors demonstrated that the MAP4K5-V569E variant emerges from an SF3B1-dependent splicing event. Mass spectrometry of the recombinant SBP-tagged MAP4K5V569E and MAP4K5V569del proteins pull-downs in transfected cell systems was used to identify the protein-protein interactions of these two MAP4K5 isoforms and propose possible functions. Together these data highlight the utility of this integrated DNA and RNA variant detection platform to detect RNA variants in cancer cells and support future analysis of RNA variant detection in cancer tissue.

Introduction

Proteogenomics platforms aim to define the expressed and mutated genome in a diseased state to better determine the mechanisms whereby signal transduction is re-wired through mutant protein signalling [1]. These approaches have been implemented in many cancer types including ovarian, colorectal, endometrial, lung, prostate and renal tissues [2–6]. Although disease-specific signalling maps can be constructed, whole-genome cancer sequencing has revealed a patient-specific cancer barcode [7,8] that complicates stratification of patients based on gene mutation status. In addition, although the vast majority of anti-cancer medicines target wild-type proteins, there are ever emerging successes in targeting mutated enzymes (kinases) with effective drug leads [9,10]. This provides the proof of

concept that drugging mutated signalling networks has therapeutic value and presents an opportunity to develop precision, personalized therapeutics based on expressed, mutant proteins.

Understanding the expression of mutant proteins in a diseased state could form a platform for the development of a range of mutation-dependent therapeutics [11]. For example, mutant neoantigen vaccines could be developed using scaffolds such as synthetic mutant proteins [12], activated dendritic cells [13], nucleic acids including RNA [14] or synthetic viral vectors [15]. Common sets of mutated neoantigens can be identified in patients with microsatellite instability cancers emerging from mutations within microsatellite regions [16]. Nevertheless, the study of mutant proteome landscapes is only in its infancy. This task is complicated because building mutant proteomes involves the

CONTACT Małgorzata Kurkowiak ✉ Malgorzata.Kurkowiak@ug.edu.pl 🏢 International Centre for Cancer Vaccine Science (ICCVS), University of Gdańsk, Kładki 24, 80-822 Gdańsk, Poland; Ted R. Hupp ✉ theodore.hupp@ug.edu.pl 🏢 International Centre for Cancer Vaccine Science (ICCVS), University of Gdańsk, Kładki 24, 80-822 Gdańsk, Poland and University of Edinburgh, Institute of Genetics and Molecular Medicine, Edinburgh Cancer Research Centre, Edinburgh, Scotland, UK,
^Current address: UMR9002 CNRS-UM, Institut De Génétique Humaine-Université De Montpellier, Gene Regulation Lab, Montpellier, France

*Joint first authorship

Supplemental data for this article can be accessed here.

integration of methodologies that link the fields of informatics and mass spectrometry with cancer biology. There are major challenges with integrating DNA sequencing, RNA sequencing, and mass spectrometric datasets [17]. Although packages/platforms for identification of mutations/variants in DNA sequences are perhaps more evolved [18], the algorithms for defining mutations with RNA sequencing datasets are very complex because the RNA mutation landscape is more pleiotropic than DNA. RNA mutations can be encoded by exons and by introns from pre-spliced RNA [19,20]. In addition, defining cancer-specific RNA edits [21], tumour-specific spliced mRNAs generating tumour-specific proteins [22–24], or even translation of UTRs provide novel mutant protein signalling landscapes [25]. As the diversity in software and computational tools tend not to be benchmarked against each other, there is as of yet no unified and validated roadmap.

In a previous report, we have benchmarked an integrated DNA and RNA variant identification software platform (CLC) to define expressed, p53-dependent single nucleotide variants (SNVs) in a human melanoma cell model [26]. The expressed SNVs detected in mRNA from the cell lines were validated using mass spectrometry in order to define the integrity of the variant identification software. This approach yielded mutant protein signal transduction maps that are enriched in either the wt-p53 or the p53-null isogenic cell model. In addition to RNA mutations that are genetically encoded, it is of interest to expand on the type of RNA variants, including RNA edits, that deviate from the genomic DNA sequence in tumour cell models. Emerging software tools are improving the detection of A-to-I RNA editing events [27,28] including developing a library of edited targets [29–32].

Here we expand on the use of the CLC integrated DNA and RNA variant detection software to define both the classic RNA edits and also multiple-nucleotide variants (MNVs) that reflect small exons not previously annotated to a reference genome. Prior to our current study, the CLC integrated DNA and RNA variant detection software has been used to identify RNA editing events in 34 protein-coding mitochondrial transcripts of four *Populus* species, a genus with a relatively small number of RNA editing sites relative to other angiosperms [33]. We were able to define sets of classical RNA edits as well as MNVs in the tumour models. We focus on validating one of these MNVs; the differential splicing of a previously non-annotated three base pair, highly conserved exon in the *MAP4K5* gene that results in the insertion of one amino acid in the protein. These data together establish the utility of this integrated DNA and RNA variant identification software to discover novel RNA variant landscapes in cell models and further highlights that understanding proteome regulation requires more accurate tools to define the RNA reference set.

Results

Defining the global RNA variant landscape in A375 melanoma cells

In a prior report [26], the CLC variant detection platform was benchmarked against *Varscan2* to identify DNA encoded expressed RNA variants in an isogenic wt-p53 A375 and p53-null A375 melanoma cell model. A total of 1468 mutated mRNA species encoded by 989 mutated genes were defined [26]. These data were then used as a reference database to define the baseline p53-dependent and p53-independent mutant proteome networks using mass spectrometry which includes over 300 mutant proteins [26]. Here we aim to expand on the RNA variant landscape from wt-p53 and p53-null A375 melanoma cells, without and with interferon treatment, that deviate from the reference genomic DNA sequence. The reason for using four biological states from highly similar genomes was to focus on common variant pathways that reflect robustness of the RNA variant type. Such RNA variants might be derived from; (i) classical A-I RNA editing events; (ii) non-canonical splicing events generating novel MNVs (exons) that have not been annotated by the reference transcriptome; or (iii) possible pseudogene expression incorrectly mapped to a reference gene but not reflecting true RNA edits (Fig. 1(a)). The RNA variants were filtered using the Fisher's exact test which was used previously to define the likelihood of true variance in RNA edits (Fig. 1(b)) [34,35].

Using the Fisher's exact test, the number of variants detected that are nonsynonymous, synonymous or non-coding are presented in Supplementary Tables 1A-G. In the four biological states used, the range of non-synonymous variants with 2 RNA variant reads or higher ranges from 3392 to 6377 (Table 1(a), i). The range of total non-synonymous variants with 10 RNA variant reads or higher ranges from 102–559 (Table 1(a), iii). The range of non-synonymous SNV variants with 10 RNA variant reads or higher ranges from 29 to 197 (Table 1(a), iii sublevel). Specific RNA variants (i.e. A-G and C-T) detected from 2–10 reads or higher identified in these four biological states are highlighted in Table 1(b) (i, ii, iii). For example, with a minimum of 10 variant RNA reads, the number of ADAR-like A-I variants (defined as A/G sequencing reads) that passes the Fisher's exact test filtering was 154 and the number of APOBEC-like C-U variants (defined as C/T sequencing reads) was 163 (Table 1(b), iii). G to A variants (Table 1(b)) might arise from APOBEC3A-dependent processes [36], whilst U-C variants (Table 1(b)) might arise from transamination and transglycosylation [37]. A calculation of the extent of classic non-synonymous RNA editing events (A-I + C-U; 720) over the total non-synonymous RNA variants from Supplementary Table 1A (3,123) is 23.05%.

Examples of specific SNV variants of interest are summarized in Table 2. One non-coding RNA variant we include (from Supplementary Table 1 F) is the miRNA, *MIR663AHG*, which has two A-G variants represented by over 1,000 reads each (Table 2; Supplementary Table 1 F). *MIR663AHG* has not been previously shown to be a target of the RNA editing machinery; however, other miRNAs have been shown to be edited by the ADAR-dependent mechanisms [38,39]. Several protein kinases are highlighted including *CDK13, CDK10, CDK11*, and *CDK12*, which exhibit A-G variants in the sequencing reads (Table 2). For example, *CDK13* exhibits 17 out of 22 variant A-G RNA reads, with a total of 43 wt-DNA reads yielding a Fisher's exact test p-value of 1.4 e-11. *CDK10* exhibits 12 out of 20 variant
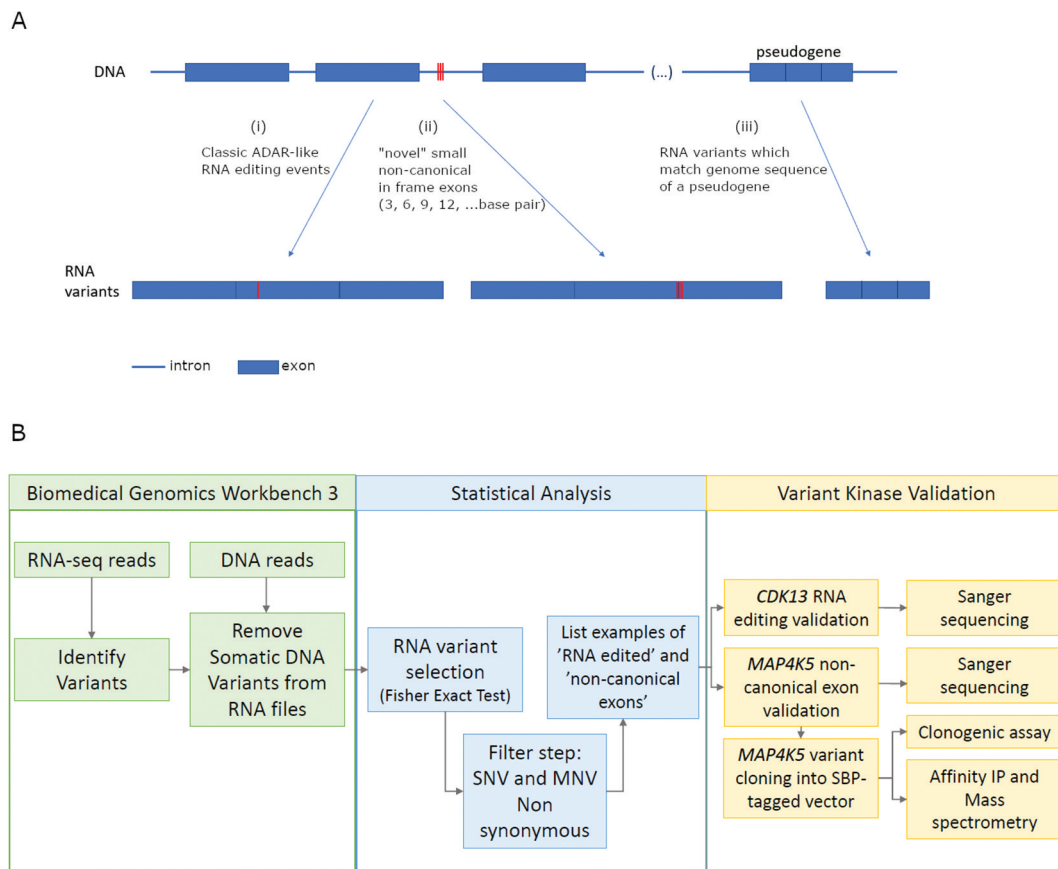
A

B

Figure 1. RNA variant detection processes.

(a) RNA variants that do not match the reference DNA can be a result of (i) classic RNA editing events such as those derived from ADAR and APOBEC activity; (ii) novel, non-canonical exons whose transcripts are not yet annotated to the reference human genome (hg19 and hg38); and (iii) potentially false RNA editing events that might arise through the expression of pseudogenes. (b) Process for annotating RNA variants that deviate from the reference genome. Left panel: Shotgun RNA sequencing reads and genomic exon-sequencing reads were imported into the *CLC* workbench. RNA reads that match somatic DNA mutations were removed leaving a file of RNA variants that do not match the reference genome or the derived transcriptome. Middle panel: The Fisher exact statistical test [34,35] was used to stratify the relative probability that the RNA variant was a true deviation from the reference genomic DNA. In the filtering steps of high probability edited RNAs, a subset was chosen that includes genes that would produce a non-synonymous mutation or variant including SNVs, MNVs, indels, and replacements (Supplementary Table 1A-E). Right panel: We chose to validate using RT-PCR and Sanger sequencing two protein kinases transcripts that represent either a classic A-I RNA edit (in *CDK13*) or a small 3 base exon (in *MAP4K5*) that was not previously annotated.

A-G RNA reads, with a total of 45 wt-DNA reads yielding a Fisher's exact p-value of 3.1 e-08. These A-G variants are most likely ADAR-dependent A-I RNA editing events. A representative nucleotide insertion event (C) in the RNA encoding *SEPT7* produces a theoretical frame shift; *SEPT7* exhibits 103 out of 146 variant C insertion RNA reads, with a total of 19 wt-DNA reads yielding a Fisher's exact p-value of 1.1 e-09. A gene with a representative MNV (*YTHDF3*) exhibits 13 out of 15 RNA variant reads, with a total of 66 wt-DNA reads yielding a Fisher's exact p-value of 2.7e-13. This representative MNV is most likely an alternate or non-canonical short exon, similar to *MAP4K5* (See below).

Summarized next are groups of genes with high-confidence, canonical A-I (shown as A-G change in the sequencing results) ADAR-like editing events (Table 3). Several genes of interest from the stress-activated signalling and cancer field emerge and we generally focus on protein kinases which form drugabble targets in oncology. We first focus on the previously identified edits in the cell cycle-dependent kinase superfamily member, *CDK13* [40]. The latter study linked this editing event to poor prognosis in hepatocellular cancer patients, but biochemical

characterization to provide a potential mechanism was not performed. The identification of this *CDK13* mRNA editing event in the A375 cell line using the *CLC* software (predicted amino acid change of Gln103Arg) forms a type of internal validation that suggests this informatics tool has value for identifying RNA variants from a reference set. In addition, to the ADAR-like editing events, we also present a list of the top genes with canonical C-U (presented as C-T changes in sequencing data) APOBEC-like editing events (Table 4). For example, *HSPD1* mRNA exhibited a variant rate of 70.31% (45 out of 64 RNA sequencing reads are C to T). In general, it is thought that APOBEC RNA editing forms a smaller landscape than ADAR, but in our analysis, C-U variants were as wide-spread as A-I variants (Table 1(b)). Future research will determine whether these C-U variants are APOBEC-dependent. We also include non-canonical RNA variants (A-T) of unknown origin that emerge from this software. A-T variants only represent a small proportion of the total, classical RNA editing events if variants are included that are only above 10% of the total RNA reads (Table 5). If the total A-T RNA variants are included from this analysis, then there are hundreds that fall below the 10%

Table 1. Different types of RNA variants that pass the Fisher's exact test. (A) Number of variant types in isogenic A375 cell lines (p53 WT and null, treated and not treated with interferon (IFN)). (B) Number of A-G and C-T variants in A375 p53-WT.

A

RNA variants that pass the Fisher's exact test		Numer of variants			
		WT-p53	WT-p53 + IFN	p53-null	p53-null + IFN
i	*Minimum 2 variant reads in the RNA*				
	Non-synonymous	6377	5738	3392	3744
	Synonymous	1340	1307	874	1013
	Non protein coding	5326	3748	3484	3617
ii	*Minimum 5 variant reads in the RNA*				
	Non-synonymous	1721	1234	251	693
	Synonymous	326	258	146	149
	Non protein coding	2751	2029	1943	1914
iii	*Minimum 10 variant reads in the RNA*				
	Non-synonymous:	559:	379:	102:	255:
	SNV	197	164	29	114
	MNV	21	22	15	18
	Indel	330	187	115	119
	Replacement	11	7	5	5
	Synonymous	105	83	62	54
	Non protein coding	1193	919	872	883

B

All RNA variants that pass the Fisher's exact test		Numer
i	*Minimum 2 variant reads in the RNA*	
	A-G	697
	G-A	1213
	C-T	1310
	T-C	714
ii	*Minimum 5 variant reads in the RNA*	
	A-G	305
	G-A	414
	C-T	453
	T-C	314
iii	*Minimum 10 variant reads in the RNA*	
	A-G	154
	G-A	150
	C-T	163
	T-C	138

threshold (Supplementary Table 1 G). These low-frequency A-T RNA variants might be either hotspots of RNA polymerase errors [41] or hotspot sequencing errors arising from the next-generation shotgun sequencing methodology [42]. Although the origin of such A-T variants are not known, in this manuscript we validate one of the high-frequency variants as being attributed to a non-annotated small exon in *MAP4K5* (Table 5).

Using the totality of these RNA variants, we evaluated the dominating gene family groups that exhibit overlapping or diverse RNA variant expression in the four biological states (Supplementary Figure 1A). In general, for transcription gene products ('Transcription, DNA-templated' and 'Regulation of Transcription'; Supplementary Figure 1A), the wt-p53 cells (untreated) exhibited similar extents of RNA variants as p53-null cells (treated with interferon). We focus here on summarized the 'DNA repair' gene family of proteins which are highest in the p53-null cells (not treated with interferon) (Supplementary Figure 1A). When these were dissected gene-by-gene, then the gene products that exhibit the elevated variant mRNA production are listed in Supplementary Figure 1B. Together, these data highlight the feasibility of using the *CLC* software to study changes in the RNA variant landscape. Below, we focus on validating two key RNA variants that represent different types of variants that can be measured; an RNA editing event in *CDK13* and a non-canonical splicing event in *MAP4K5*.

Sanger-sequencing validation of the dominant RNA editing events in the *CDK13* mRNA

The genome browser view of the region in *CDK13* which is edited at the RNA level is shown in Fig. 2(a). *CDK13* shows A-to-I editing within exon 1 in A375 melanoma cells as well as SiHa cervical carcinoma cells. These data, along with prior reports of *CDK13* RNA editing [40,43], suggests that editing at this position is widespread. The dominant edit would change a glutamine to an arginine (Q to R) in codon 103 if the mRNA were translated. This high-level editing in *CDK13* transcript suggests a functional role for the predominant edited variant of the gene product [44]. In addition, we also detect the A-G variant nearby codon 103, the K96R variant, although the reads are substantially reduced and they both exist on the same transcript (i.e. these two RNA editing events are not apparently mutually exclusive on the same transcript [40] (Fig. 2(a)).

Despite the fact that *CDK13* has been previously shown to be edited at codon 103 [40], we validated this target further using Sanger sequencing since it is the most abundant editing event. Due to the GC-rich nature of *CDK13*, nested PCR primers were first used to amplify the *CDK13* fragment from cDNA libraries (Fig. 2(b)). Using these optimized primers, we could detect edited (I, recognized as G) and non-edited (A) species in a range of conditions as defined by the ratio of G and A sequencing reads (Fig. 2(c)), We were unable to detect changes in the G/A ratio in this cell model in response to UV irradiation (Fig. 2(c) i), splicing inhibitor treatment (Fig. 2(c) ii), or the replacement of cells in serum-free media (Fig. 2(c) iii). However, the long-term incubation of cells with an siRNA control reduced the G-A ratio so that the non-edited sequence predominated after 4 days of incubation (Fig. 2(d)ii vs Fig. 2(d))i. These latter data suggest that editing of *CDK13* can be regulated by signalling changes. Future work will define the function of the edited version of *CDK13* gene product and how this editing event can be regulated.

Sanger validation of the minor RNA editing event in relation to the major RNA editing event was also shown in Fig. 2 (e-g); the K96R variant exhibits less editing that the Q103R variant based on peak intensities of the G or A sequencing reads (Fig. 2(e) vs 2(f)), which is consistent with the number of reads in the *CLC* browser (Fig. 2(a) and Supplementary Table 1A). There are seven other non-synonymous RNA

Table 2. Examples of specific single nucleotide variants (SNV) of interest. SNVs in miRNA are highlighted in green, while SNVs in kinases are highlighted in yellow.

Chr	Region	Type	Ref	Allele	Count (RNA)	Coverage (RNA)	Frequency (RNA) [%]	Gene	AA change	Non-synonymous mutation	Count (DNA)	Coverage (DNA)	Fisher P-value
20	26188680	SNV	A	G	2097	2097	100	MIR663AHG		-	0	13	3.93273499519197e-34
20	26188660	SNV	A	G	1643	1644	99.94	MIR663AHG		-	0	13	1.2873922740671e-31
2	102483777	SNV	G	A	14	16	87.5	MAP4K4		No	0	37	4.9917222259838666e-11
7	39990548	SNV	A	G	17	22	77.27	CDK13	ENSP00000181839: p.Gln103Arg	Yes	0	43	1.40982069224812e-11
16	89754046	SNV	A	G	12	20	60	CDK10	ENSP00000426264: p.Gln46Arg	Yes	0	45	3.1275056379363603e-08
1	1653093	SNV	A	G	7	24	29.17	CDK11A, RP1-283E3.8	ENSP00000348529: p.Cys23Arg	Yes	0	67	4.2760615206527 5e-05
17	37687369	SNV	G	A	4	54	7.41	CDK12	ENSP00000398880: p.Gly1425Arg	Yes	0	146	0.0048890970 7745
17	60600495	SNV	G	C	10	97	10.31	TLK2	ENSP00000275780: p.Gly75Ala	Yes	0	77	0.00253844779949121
8	64087986..64087988	MNV	TGA	CAC	13	15	86.67	YTHDF3	ENSP00000473496: p.Leu49_Thr50delinsProPro	Yes	0	66	2.79714370445 54e-13
7	35840881^35840882	Insertion	-	C	103	146	70.55	SEPT7	ENSP00000344868: p.Gly21fs	Yes	0	19	1.128428401 56902e-09
12	69653974^69653975	Insertion	-	AAG	41	289	14.19	CPSF6	ENSP00000391774: p.Arg491_Glu492insArg	Yes	0	49	0.001590436 60642 8532
9	77703640^77703641	Insertion	-	AGGTGAGGG	4	69	5.8	OSTF1	ENSP00000340836: p.Gly12_Gln13insGluGlyGly	Yes	0	96	0.0290366380 146419
4	83350581...83350583	Replacement	CGG	AT	14	62	22.58	HNRNPDL	ENSP00000295470: p.Arg88fs	Yes	0	93	1.00475202732196e-06

Table 3. Summarized groups of genes with high-confidence A-G variants.

Chr	Region	Type	Presence in REDIportal	Ref	Allele	Count (RNA)	Coverage (RNA)	Frequency (RNA) [%]	Gene	AA change	Non-synonymous mutation	Count (DNA)	Coverage (DNA)	Fisher P-value
19	14,883,288	SNV	No	A	G	50	51	98.04	EMR2	ENSP00000319883:p.Leu74Pro	Yes	0	54	1.82112894190718e-29
1	149,400,160	SNV	No	A	G	49	74	66.22	RP5-998N21.10, RP5-998N21.7, HIST2H3PS2	ENSP00000476960:p.Val128Ala	Yes	0	57	1.22483931860094e-17
1	225,974,614	SNV	Yes (also in ATLAS, RADAR, DANRED)	A	G	43	96	44.79	SRP9	ENSP00000355804:p.Ile64Met	Yes	0	7	0.03957993607879
6	32,557,461	SNV	No	A	G	33	66	50	HLA-DRB1	ENSP00000353099:p.Met20Thr	Yes	0	20	1.26361088993366e-05
17	7,479,846	SNV	No	A	G	32	35	91.43	SENP3-EIF4A1, EIF4A1, SNORA67	ENSP00000293831:p.Gln117Arg	Yes	0	32	5.71510505664653e-16
12	58,223,322	SNV	No	A	G	20	102	19.61	CTDSP2	ENSP00000381148:p.Leu41Pro	Yes	0	48	0.000435700500593407
7	39,990,548	SNV	Yes (also in ATLAS, RADAR)	A	G	17	22	77.27	CDK13	ENSP00000181839:p.Gln103Arg	Yes	0	43	1.40982069224812e-11
12	58,223,310	SNV	No	A	G	17	100	17	CTDSP2	ENSP00000381148:p.Phe45Ser	Yes	0	45	0.00155609502793716
7	56,183,824	SNV	No	A	G	15	16	93.75	NUPR1L	ENSP00000455442:p.Trp62Arg	Yes	0	75	2.91788201938187e-16
13	42,887,184	SNV	No	A	G	15	32	46.88	AKAP11	ENSP00000002301:p.Asn1759Asp	Yes	0	73	1.01637496378703e-09
3	155,643,079	SNV	No	A	G	14	217	6.45	GMPS	ENSP00000419851:p.Gln495Arg	Yes	0	109	0.006340656844091
16	89,754,046	SNV	No	A	G	12	20	60	CDK10	ENSP00000426264:p.Gln46Arg	Yes	0	45	3.12750563793603e-08
3	37,067,261	SNV	No	A	G	12	146	8.22	MLH1	ENSP00000231790:p.Gln391Arg	Yes	0	47	0.0411804905311472
7	56,183,781	SNV	No	A	G	11	15	73.33	NUPR1L	ENSP00000455442:p.Leu76Pro	Yes	0	46	3.26481485245141e-09
12	123,879,645	SNV	No	A	G	11	194	5.67	SETD8	ENSP00000384629:p.Gln114Arg	Yes	0	144	0.00309856873055919
12	58,221,358	SNV	No	A	G	11	114	9.65	CTDSP2	ENSP00000381148:p.Cys77Arg	Yes	0	86	0.00276332413384754
4	41,668,668	SNV	No	A	G	10	152	6.58	LIMCH1	ENSP00000316891:p.Asn743Asp	Yes	0	86	0.0152486973238669
18	669,140	SNV	No	A	G	10	145	6.9	TYMS	ENSP00000315644:p.Arg175Gly	Yes	0	91	0.00775777613161119
16	74,516,925	SNV	No	A	G	10	145	6.9	GLG1	ENSP00000405984:p.Trp557Arg	Yes	0	105	0.00588521683893524
12	32,891,213	SNV	No	A	G	9	114	7.89	DNM1L	ENSP00000415131:p.Arg538Gly	Yes	0	62	0.0277227730166222
12	121,855,472	SNV	No	A	G	9	130	6.92	RNF34	ENSP00000355137:p.Asn131Asp	Yes	0	94	0.0112941289917348
17	40,714,905	SNV	No	A	G	9	148	6.08	COASY	ENSP00000377406:p.Asn89Asp	Yes	0	118	0.00513004249124648
X	55,103,084	SNV	No	A	G	8	8	100	PAGE2B	ENSP00000364110:p.Glu56Gly	Yes	0	65	7.43929938132365e-11
2	27,464,094	SNV	No	A	G	8	148	5.41	CAD	ENSP00000264705:p.Gln1936Arg	Yes	0	167	0.00214257228352634
12	133,588,015	SNV	No	A	G	7	28	25	ZNF26	ENSP00000437420:p.Lys497Arg	Yes	0	78	4.85858311809867e-05
1	1,653,093	SNV	No	A	G	7	24	29.17	CDK11A, RP1-283E3.8	ENSP00000348529:p.Cys23Arg	Yes	0	67	4.27606152065275e-05
11	111,631,628	SNV	No	A	G	7	91	7.69	PPP2R1B	ENSP00000437193:p.Trp152Arg	Yes	0	56	0.00443114605110571

(Continued)

Table 3. (Continued).

Chr	Region	Type	Presence in REDIportal	Ref	Allele	Count (RNA)	Coverage (RNA)	Frequency (RNA) [%]	Gene	AA change	Non-synonymous mutation	Count (DNA)	Coverage (DNA)	Fisher P-value
1	32,622,469	SNV	No	A	G	7	102	6.86	KPNA6	ENSP00000362728:p.Asn52Asp	Yes	0	70	0.0423127006073164
19	21,301,076	SNV	Yes (also in ATLAS, RADAR)	A	G	7	38	18.42	ZNF714	ENSP00000472368:p.Lys536Glu	Yes	0	29	0.0163066351077906
5	68,390,093	SNV	No	A	G	7	79	8.86	SLC30A5	ENSP00000379836:p.Lys4Arg	Yes	0	96	0.00328207326350711
15	56,134,222	SNV	No	A	G	7	80	8.75	NEDD4	ENSP00000424827:p.Leu1002Ser	Yes	0	105	0.00242154665846764
10	127,530,479	SNV	No	A	G	7	70	10	BCCIP, DHX32	ENSP00000284690:p.Leu459Ser	Yes	0	98	0.0018153179968
1	171,509,868	SNV	No	A	G	7	61	11.48	PRRC2C	ENSP00000356716:p.Lys1088Arg	Yes	0	105	0.000719618475390669
1	197,111,808	SNV	No	A	G	7	129	5.43	ASPM	ENSP00000356379:p.Val525Ala	Yes	0	244	0.000530692684709058
11	120,335,968	SNV	No	A	G	7	81	8.64	ARHGEF12	ENSP00000380942:p.Lys879Arg	Yes	0	173	0.000279281870195161
5	176,942,758	SNV	No	A	G	6	100	6	DDX41	ENSP00000422753:p.Tyr167His	Yes	0	72	0.0409477773535999
3	182,591,758	SNV	No	A	G	6	103	5.83	ATP11B	ENSP00000321195:p.Lys736Arg	Yes	0	87	0.0320681506824627
5	76,732,198	SNV	No	A	G	6	104	5.77	WDR41	ENSP00000296679:p.Phe372Ser	Yes	0	92	0.0306013550098557
19	45,153,108	SNV	No	A	G	6	61	9.84	PVR	ENSP00000402060:p.Gln152Arg	Yes	0	62	0.0130662556117345
19	49,303,354	SNV	No	A	G	6	99	6.06	BCAT2	ENSP00000385161:p.Phe99Leu	Yes	0	127	0.00647390928657926
4	38,879,797	SNV	No	A	G	5	66	7.58	FAM114A1	ENSP00000351740:p.His33Arg	Yes	0	75	0.0206732093679803
17	25,633,838	SNV	No	A	G	5	79	6.33	WSB1	ENSP00000262394:p.Gln214Arg	Yes	0	94	0.0185003944601614
X	134,986,647	SNV	No	A	G	5	31	16.13	SAGE1	ENSP00000445959:p.Thr78Ala	Yes	0	38	0.0151186371364254
3	50,138,033	SNV	No	A	G	5	56	8.93	RBM5	ENSP00000343054:p.Asn160Asp	Yes	0	80	0.0106125718158434
6	74,446,113	SNV	No	A	G	5	86	5.81	CD109	ENSP00000388062:p.Lys172Arg	Yes	0	132	0.00889276107366033
19	53,856,722	SNV	No	A	G	5	16	31.25	ZNF845	ENSP00000388311:p.Ile932Val	Yes	0	29	0.00357517317245054
19	53,385,141	SNV	No	A	G	5	39	12.82	ZNF320	ENSP00000473091:p.Tyr80His	Yes	0	127	0.000582481281575573

Table 4. Summarized groups of genes with high-confidence C-T variants.

Chr	Region	Type	Ref	Allele	Count (RNA)	Coverage (RNA)	Frequency (RNA) [%]	Gene	AA change	Non-synonymous mutation	Count (DNA)	Coverage (DNA)	Fisher P-value
2	198,361,863	SNV	C	T	45	64	70.31	HSPD1	ENSP00000441296:p.Gly143Asp	Yes	0	59	1.25874384373074e-18
20	17,639,820	SNV	C	T	38	443	8.58	RRBP1	ENSP00000367044:p.Ala445Thr	Yes	0	59	0.0149732630163734
20	17,639,790	SNV	C	T	31	410	7.56	RRBP1	ENSP00000246043:p.Ala455Thr	Yes	0	60	0.0226710686326192
1	241,767,881	SNV	C	T	28	37	75.68	OPN3	ENSP00000355512:p.Gly125Glu	Yes	0	57	1.89876937313551e-16
7	151,933,014	SNV	C	T	28	87	32.18	KMT2C	ENSP00000347325:p.Arg886His	Yes	0	20	0.00149079919815577
7	151,932,997	SNV	C	T	28	92	30.43	KMT2C	ENSP00000262189:p.Gly892Arg	Yes	0	19	0.00304242853041439
12	58,223,346	SNV	C	T	27	93	29.03	CTDSP2	ENSP00000448951:p.Arg33His	Yes	0	53	1.14233434634688e-06
20	3,765,524	SNV	C	T	22	415	5.30	CENPB	ENSP00000369075:p.Gly536Asp	Yes	0	81	0.0343572449619834
9	138,758,382	SNV	C	T	22	63	34.92	CAMSAP1	ENSP00000374183:p.Val196Ile	Yes	0	23	0.000490132112142799
1	16,890,676	SNV	C	T	20	57	35.09	NBPF1	ENSP00000474456:p.Ser1061Asn	Yes	0	11	0.0261283744915462
12	58,223,305	SNV	C	T	17	97	17.53	CTDSP2	ENSP00000381148:p.Ala47Thr	Yes	0	44	0.00150595341736571
17	16,097,786	SNV	C	T	16	93	17.20	NCOR1	ENSP00000387727:p.Arg33His	Yes	0	46	0.00134445382694989
12	58,218,102	SNV	C	T	15	104	14.42	CTDSP2	ENSP00000381148:p.Val138Met	Yes	0	49	0.00283758574140028
19	6,741,282	SNV	C	T	15	127	11.81	TRIP10	ENSP00000469360:p.Pro63Ser	Yes	0	68	0.00145574724301353
13	36,909,801	SNV	C	T	14	189	7.41	SPG20	ENSP00000414147:p.Gly56Glu	Yes	0	72	0.0130234368145754
11	102,953,493	SNV	C	T	14	203	6.90	DCUN1D5	ENSP00000260247:p.Gly109Arg	Yes	0	68	0.0243485205933667
4	6,711,266	SNV	C	T	13	139	9.35	MRFAP1L1	ENSP00000318154:p.Val31Ile	Yes	0	55	0.0212807038888119
X	102,508,880	SNV	C	T	13	113	11.50	TCEAL8	ENSP00000353093:p.Gly10Arg	Yes	0	75	0.00188230611527362
11	104,820,408	SNV	C	T	12	159	7.55	CASP4	ENSP00000388566:p.Gly215Arg	Yes	0	60	0.0395485736622814
X	64,754,507	SNV	C	T	12	165	7.27	LAS1L	ENSP00000473471:p.Gly30Glu	Yes	0	190	8.14320230043981e-05
7	158,445,167	SNV	C	T	12	61	19.67	NCAPG2	ENSP00000388326:p.Gly918Arg	Yes	0	41	0.00143797023710544
11	71,715,090	SNV	C	T	12	114	10.53	NUMA1	ENSP00000260051:p.Gly924Glu	Yes	0	85	0.00140855481829364
16	14,859,247	SNV	C	T	12	12	100	NPIPA2	ENSP00000432029:p.Arg363Trp	Yes	0	3	0.0021978021978022
7	140,154,980	SNV	C	T	11	189	5.82	MKRN1	ENSP00000255977:p.Gly384Glu	Yes	0	69	0.0397674501449123
16	68,191,775	SNV	C	T	11	31	35.48	NFATC3	ENSP00000454451:p.Pro36Leu	Yes	0	48	9.36951954107799e-06
7	127,973,362	SNV	C	T	10	117	8.55	RBM28	ENSP00000223073:p.Gly334Glu	Yes	0	89	0.0055518653895707
10	88,259,939	SNV	C	T	10	129	7.75	WAPAL	ENSP00000298767:p.Arg354Gln	Yes	0	189	9.7351812610261e-05
5	143,543,718	SNV	C	T	10	145	6.90	YIPF5	ENSP00000397704:p.Gly129Asp	Yes	0	88	0.0148132140219
11	47,505,962	SNV	C	T	10	163	6.14	CELF1	ENSP00000436864:p.Gly169Arg	Yes	0	84	0.0175118335014926
15	58,913,707	SNV	C	T	10	163	6.14	ADAM10	ENSP00000260408:p.Gly492Arg	Yes	0	187	0.000412333791144602
20	377,219	SNV	C	T	10	174	5.75	TRIB3	ENSP00000415416:p.Pro348Leu	Yes	0	180	0.000719031949974406
12	76,804,402	SNV	C	T	10	100	10	OSBPL8	ENSP00000261183:p.Gly77Glu	Yes	0	49	0.0308234990088144
X	122,766,729	SNV	C	T	9	101	8.91	THOC2	ENSP00000347959:p.Gly767Arg	Yes	0	76	0.0108213950102794
12	58,217,833	SNV	C	T	9	103	8.74	CTDSP2	ENSP00000381148:p.Val182 Met	Yes	0	88	0.0040018560048213
10	101,997,834	SNV	C	T	9	143	6.29	CWF19L1	ENSP00000326411:p.Gly400Glu	Yes	0	71	0.0312686936937371
10	97,447,031	SNV	C	T	9	65	13.85	TCTN3	ENSP00000265993:p.Gly255Arg	Yes	0	78	0.000599795019212057
1	222,732,066	SNV	C	T	8	88	9.09	TAF1A	ENSP00000327072:p.Gly430Glu	Yes	0	53	0.0248771943534809
3	111,304,184	SNV	C	T	8	118	6.78	CD96	ENSP00000283285:p.Pro272Ser	Yes	0	95	0.00927347172924355
3	55,844,280	SNV	C	T	8	129	6.20	SMEK2	ENSP00000272313:p.Gly48Arg	Yes	0	123	0.00708427312593516
3	112,280,348	SNV	C	T	8	151	5.30	ATG3	ENSP00000420259:p.Gly10Arg	Yes	0	98	0.0238356851191133
15	28,391,439	SNV	C	T	8	42	19.05	HERC2	ENSP00000261609:p.Arg3651His	Yes	0	89	6.82328197605092e-05
13	23,928,934	SNV	C	T	8	64	12.5	SACS	ENSP00000371735:p.Gly606Glu	Yes	0	89	0.000715857797288114
5	10,417,482	SNV	C	T	8	73	10.96	MARCH6	ENSP00000274140:p.Pro750Leu	Yes	0	43	0.0250897173719705
13	37,619,420	SNV	C	T	8	79	10.13	SUPT20H	ENSP00000419754:p.Gly87Arg	Yes	0	91	0.00178233562399633
11	57,076,547	SNV	C	T	8	86	8.14	TNKS1BP1	ENSP00000350990:p.Gly1213Glu	Yes	0	96	0.00460131172536528
15	56,134,226	SNV	C	T	7	88	7.96	NEDD4	ENSP00000424827:p.Gly1001Arg	Yes	0	107	0.00332727207990333
17	28,599,835	SNV	C	T	7	89	7.87	BLMH	ENSP00000261714:p.Gly295Glu	Yes	0	130	0.001583509804342
6	155,581,413	SNV	C	T	7	103	6.80	TFB1M	ENSP00000356134:p.Gly263Glu	Yes	0	59	0.00486704448703024
10	101,915,939	SNV	C	T	7	103	6.80	ERLIN1	ENSP00000410964:p.Met236Ile	Yes	0	115	0.00470385687118436
15	101,718,348	SNV	C	T	7	107	6.54	CHSY1	ENSP00000254190:p.Gly552Arg	Yes	0	95	0.0151457077745357
9	21,333,906	SNV	C	T	7	108	6.48	KLHL9	ENSP00000351933:p.Arg318Gln	Yes	0	174	0.00106812510654314
9	125,585,299	SNV	C	T	7	109	6.42	PDCL	ENSP00000259467:p.Gly117Glu	Yes	0	67	0.0451669622310842
6	86,246,557	SNV	C	T	7	109	6.42	SNX14	ENSP00000313121:p.Gly521Arg	Yes	0	135	0.00318115735029092
1	28,857,124	SNV	C	T	7	12	58.33	RCC1	ENSP00000362937:p.Pro55Ser	Yes	0	99	2.33190502692993e-08
17	57,775,101	SNV	C	T	7	124	5.65	PTRH2	ENSP00000387180:p.Gly81Glu	Yes	0	160	0.00274229316012287

(Continued)

Table 4. (Continued).

Chr	Region	Type	Ref	Allele	Count (RNA)	Coverage (RNA)	Frequency (RNA) [%]	Gene	AA change	Non-synonymous mutation	Count (DNA)	Coverage (DNA)	Fisher P-value
1	246,729,361	SNV	C	T	7	130	5.38	TFB2M	ENSP00000355471:p.Gly27Glu	Yes	0	80	0.0457812335157669
1	200,522,681	SNV	C	T	7	132	5.30	KIF14	ENSP00000356319:p.Trp1594*	Yes	0	105	0.0184523225925478
20	33,969,744	SNV	C	T	7	132	5.30	UQCC1	ENSP00000398531:p.Gly118Arg	Yes	0	111	0.0166866857610569
16	11,827,898	SNV	C	T	7	42	16.67	TXNDC11	ENSP00000283033:p.Gly170Glu	Yes	0	40	0.0120002120619595
10	51,123,969	SNV	C	T	7	47	14.89	PARG	ENSP00000384408:p.Trp92*	Yes	0	45	0.0123589350527848
9	86,495,276	SNV	C	T	7	10	70	KIF27	ENSP00000297814:p.Arg860Gln	Yes	0	193	4.7266541842133 7e-11
5	98,221,290	SNV	C	T	6	61	9.84	CHD1	ENSP00000284049:p.Gly854Arg	Yes	0	66	0.0107412065230634
5	132,227,887	SNV	C	T	6	61	9.84	AFF4	ENSP00000265343:p.Gly869Glu	Yes	0	72	0.00809811748698665
5	70,650,500	SNV	C	T	6	61	9.84	LRRC40	ENSP00000359990:p.Gly169Arg	Yes	0	85	0.00458033379491009
3	23,934,608	SNV	C	T	6	63	9.52	NKIRAS1	ENSP00000396063:p.Gly186Glu	Yes	0	72	0.0090451243666907
16	68,380,182	SNV	C	T	6	65	9.23	PRMT7	ENSP00000454776:p.Thr397Ile	Yes	0	46	0.0406165295081866
10	119,043,909	SNV	C	T	6	66	9.09	PDZD8	ENSP00000334642:p.Gly779Arg	Yes	0	89	0.00520287864187153
12	116,453,025	SNV	C	T	6	68	8.82	MED13L	ENSP00000281928:p.Gly355Glu	Yes	0	64	0.0281686928835132
12	58,217,838	SNV	C	T	6	69	8.70	ARID4B	ENSP00000264183:p.Gly582Glu	Yes	0	180	0.00084796464701192
8	59,502,088	SNV	C	T	6	80	7.5	NSMAF	ENSP00000411012:p.Gly743Arg	Yes	0	108	0.00531158603809772
5	891,357	SNV	C	T	6	84	7.14	BRD9	ENSP00000419765:p.Gly105Arg	Yes	0	66	0.0347863388868885
21	19,704,462	SNV	C	T	6	84	7.14	TMPRSS15	ENSP00000284885:p.Trp531*	Yes	0	72	0.0309839909312811
12	76,767,173	SNV	C	T	6	95	6.32	OSBPL8	ENSP00000261183:p.Gly623Glu	Yes	0	126	0.00575144655643513
12	89,992,916	SNV	C	T	6	104	5.77	CTDSP2	ENSP00000381148:p.Cys180Tyr	Yes	0	87	0.0324661227483437
12	112,743,983	SNV	C	T	6	108	5.56	ATP2B1	ENSP00000261173:p.Gly1110Asp	Yes	0	105	0.0291612251146297
9	117,844,992	SNV	C	T	6	113	5.31	TNC	ENSP00000443478:p.Trp742*	Yes	0	92	0.03383532955201289
1	6,659,126	SNV	C	T	6	116	5.17	KLHL21	ENSP00000366886:p.Gly470Arg	Yes	0	81	0.00437866381921444
10	88,260,312	SNV	C	T	6	119	5.04	WAPAL	ENSP00000298767:p.Gly230Arg	Yes	0	149	0.00713415055 2398
1	31,501,660	SNV	C	T	6	119	5.04	PUM1	ENSP00000362846:p.Gly175Arg	Yes	0	154	0.00637857742112903
19	5,679,241	SNV	C	T	6	13	46.15	C19orf70	ENSP00000465739:p.Gly109Glu	Yes	0	19	0.00189368434342391
12	112,743,983	SNV	C	T	6	21	28.57	HECTD4	ENSP00000449784:p.Ser263Asn	Yes	0	62	0.00014376582334118
13	42,352,178	SNV	C	T	6	44	13.64	VWA8	ENSP00000281496:p.Met764Ile	Yes	0	66	0.00329577076425277
5	176,025,977	SNV	C	T	6	46	13.04	GPRIN1	ENSP00000305839:p.Gly287Arg	Yes	0	54	0.00785877242074 2579
10	75,888,943	SNV	C	T	6	53	11.32	AP3M1	ENSP00000361831:p.Trp242*	Yes	0	37	0.0040606601229 3993
X	76,937,786	SNV	C	T	6	55	10.91	ATRX	ENSP00000362441:p.Gly988Arg	Yes	0	53	0.0271469474581715
11	66,473,160	SNV	C	T	6	57	10.53	SPTBN2	ENSP00000433593:p.Gly601Glu	Yes	0	90	0.00287126633204993
12	112,744,038	SNV	C	T	6	20	30	HECTD4	ENSP00000366783:p.Val245Met	Yes	0	46	0.00042659614 3148232
X	76,854,946	SNV	C	T	6	50	12	ATRX	ENSP00000362441:p.Gly1964Arg	Yes	0	103	0.00098497297822 4279
14	31,598,133	SNV	C	T	6	60	10	HECTD1	ENSP00000450697:p.Gly1482Arg	Yes	0	109	0.00169242132906296
14	45,711,445	SNV	C	T	6	75	8	MIS18BP1	ENSP00000309790:p.Gly312Glu	Yes	0	90	0.00807641260466534
16	2,983,473	SNV	C	T	5	51	9.80	FLYWCH1	ENSP00000253928:p.Pro380Leu	Yes	0	94	0.00471521355544559
3	25,773,866	SNV	C	T	5	52	9.62	NGLY1	ENSP00000280700:p.Gly457Arg	Yes	0	73	0.0110815071465415
3	25,761,636	SNV	C	T	5	52	9.62	NGLY1	ENSP00000280700:p.Trp553*	Yes	0	78	0.0090795336294329
5	33,616,115	SNV	C	T	5	56	8.93	ADAMTS12	ENSP00000422554:p.Gly736Arg	Yes	0	175	0.00077279276324 2569
8	67,547,235	SNV	C	T	5	56	8.93	VCPIP1	ENSP00000309031:p.Gly1057Glu	Yes	0	182	0.00062618500 5446896
13	23,910,175	SNV	C	T	5	57	8.77	SACS	ENSP00000371729:p.Gly2614Arg	Yes	0	147	0.00149414096748599
4	140,468,098	SNV	C	T	5	60	8.33	SETD7	ENSP00000427300:p.Gly49Glu	Yes	0	67	0.0214824130461269
15	35,273,591	SNV	C	T	5	61	8.20	ZNF770	ENSP00000348673:p.Gly682Glu	Yes	0	76	0.0159252501523542
17	37,565,419	SNV	C	T	5	64	7.81	MED1	ENSP00000300651:p.Gly1019Arg	Yes	0	76	0.0182857130577 95
2	179,095,556	SNV	C	T	5	66	7.58	ABL2	ENSP00000427562:p.Gly215Arg	Yes	0	69	0.0257777100835717
2	219,529,098	SNV	C	T	5	68	7.35	RNF25	ENSP00000295704:p.Gly321Glu	Yes	0	79	0.0195201998299625
4	187,557,906	SNV	C	T	5	68	7.35	FAT1	ENSP00000406229:p.Glu1269Lys	Yes	0	169	0.00174550457677628
X	67,742,699	SNV	C	T	5	69	7.25	YIPF6	ENSP00000417573:p.Pro178Ser	Yes	0	74	0.0242049264822056
4	39,304,691	SNV	C	T	5	83	6.02	RFC1	ENSP00000371321:p.Gly732Arg	Yes	0	132	0.00794774435001324
2	9,645,403	SNV	C	T	5	84	5.95	ADAM17	ENSP00000309968:p.Gly479Glu	Yes	0	142	0.00656975877421133
5	43,299,016	SNV	C	T	5	85	5.88	HMGCS1	ENSP00000322706:p.Gly18Arg	Yes	0	165	0.00419616037612641
10	104,853,028	SNV	C	T	5	90	5.56	NT5C2	ENSP00000339479:p.Gly343Arg	Yes	0	145	0.00768067081535 03
7	152,007,061	SNV	C	T	5	90	5.56	KMT2C	ENSP00000453752:p.Gly280Glu	Yes	0	226	0.00172785254115418
13	77,581,337	SNV	C	T	5	92	5.44	FBXL3	ENSP00000347834:p.Trp410*	Yes	0	96	0.0265119978875076
16	75,448,509	SNV	C	T	5	92	5.44	CFDP1, RP11-77K12.1	ENSP00000457654:p.Gly107Glu	Yes	0	112	0.0175485315174974

(Continued)

Table 4. (Continued).

Chr	Region	Type	Ref	Allele	Count (RNA)	Coverage (RNA)	Frequency (RNA) [%]	Gene	AA change	Non-synonymous mutation	Count (DNA)	Coverage (DNA)	Fisher P-value
17	80,543,923	SNV	C	T	5	94	5.32	FOXK2	ENSP00000335677:p.His475Tyr	Yes	0	152	0.0076167530767957
13	108,863,054	SNV	C	T	5	30	16.67	LIG4	ENSP00000349393:p.Arg188Gln	Yes	0	118	0.000257841011265669
2	220,467,278	SNV	C	T	5	36	13.89	STK11IP	ENSP00000295641:p.Thr187Ile	Yes	0	56	0.00766600278080494
2	220,467,196	SNV	C	T	5	37	13.51	STK11IP	ENSP00000295641:p.Leu160Phe	Yes	0	40	0.0220620043258832
1	231,081,143	SNV	C	T	5	39	12.82	TTC13	ENSP00000355621:p.Gly191Arg	Yes	0	55	0.0104890931335979
2	64,144,056	SNV	C	T	5	45	11.11	VPS54	ENSP00000272322:p.Gly736Arg	Yes	0	114	0.00153729294019848
6	116,966,945	SNV	C	T	5	46	10.87	ZUFSP	ENSP00000357565:p.Gly541Arg	Yes	0	76	0.0066127994558258
10	73,912,705	SNV	C	T	5	47	10.64	ASCC1	ENSP00000339404:p.Gly251Glu	Yes	0	66	0.010928521598832
13	60,565,350	SNV	C	T	5	20	25	DIAPH3	ENSP00000383178:p.Gly435Arg	Yes	0	78	0.00022829926004181

Table 5. A-T variants observed.

| Chr	Region	Type	Ref	Allele	Count (RNA)	Coverage (RNA)	Frequency (RNA) [%]	Gene	AA change	Non-synonymous mutation	Count (DNA)	Coverage (DNA)	Fisher P-value
14	50,904,729	SNV	A	T	22	23	95.65	MAP4K5	ENSP00000013125:p.Val569Glu	Yes	0	22	5.82989055086052e-12
12	54,740,397	SNV	A	T	8	9	88.89	COPZ1	ENSP00000449341:p.Thr107Ser	Yes	0	47	6.3358236816299e-09
14	24,684,868	SNV	A	T	37	43	86.05	MDP1, NEDD8-MDP1	ENSP00000474249:p.Val63Glu	Yes	0	75	1.03839186240207e-24
1	120,263,791	SNV	A	T	5	6	83.33	PHGDH	ENSP00000358415:p.Gln12Leu	Yes	0	89	1.0355447454656e-07
11	33,733,049	SNV	A	T	11	14	78.57	CD59	ENSP00000436737:p.Leu58Met	Yes	0	13	3.39011780659378e-05
1	10,386,409	SNV	A	T	5	7	71.43	KIF1B	ENSP00000366290:p.Leu972Phe	Yes	0	51	4.58303543602999e-06
19	40,325,306	SNV	A	T	5	9	55.56	FBL	ENSP00000472419:p.*230Arg	Yes	0	82	2.70941766486129e-06
10	95,157,002	SNV	A	T	5	9	55.56	MYOF	ENSP00000360544:p.*446Arg	Yes	0	9	0.0294117647058823
2	70,528,734	SNV	A	T	6	11	54.55	FAM136A, AC022201.5	ENSP00000391468:p.Val30Glu	Yes	0	144	2.64556738491617e-08
8	144,900,361	SNV	A	T	7	16	43.75	PUF60	ENSP00000432091:p.Val66Glu	Yes	0	147	2.149088100266e-08
9	75,769,255	SNV	A	T	17	52	32.69	ANXA1	ENSP00000412489:p.Arg6Trp	Yes	0	54	1.12413403199044e-06
3	75,832,479	SNV	A	T	9	41	21.95	ZNF717	ENSP00000417902:p.Leu12Gln	Yes	0	25	0.010718368892246
11	244,159	SNV	A	T	3	14	21.43	PSMD13	ENSP00000396937:p.Arg72Trp	Yes	0	23	0.0468468468468469
12	58,221,366	SNV	A	T	23	111	20.72	CTDSP2	ENSP00000381148:p.Leu74Gln	Yes	0	88	5.2180189878547e-07
9	78,790,200	SNV	A	T	11	56	19.64	PCSK5	ENSP00000365958:p.Glu685Asp	Yes	0	27	0.0135143112551926
8	146,016,146	SNV	A	T	7	36	19.44	RPL8	ENSP00000433703:p.*168Arg	Yes	0	74	0.0002622325862231386
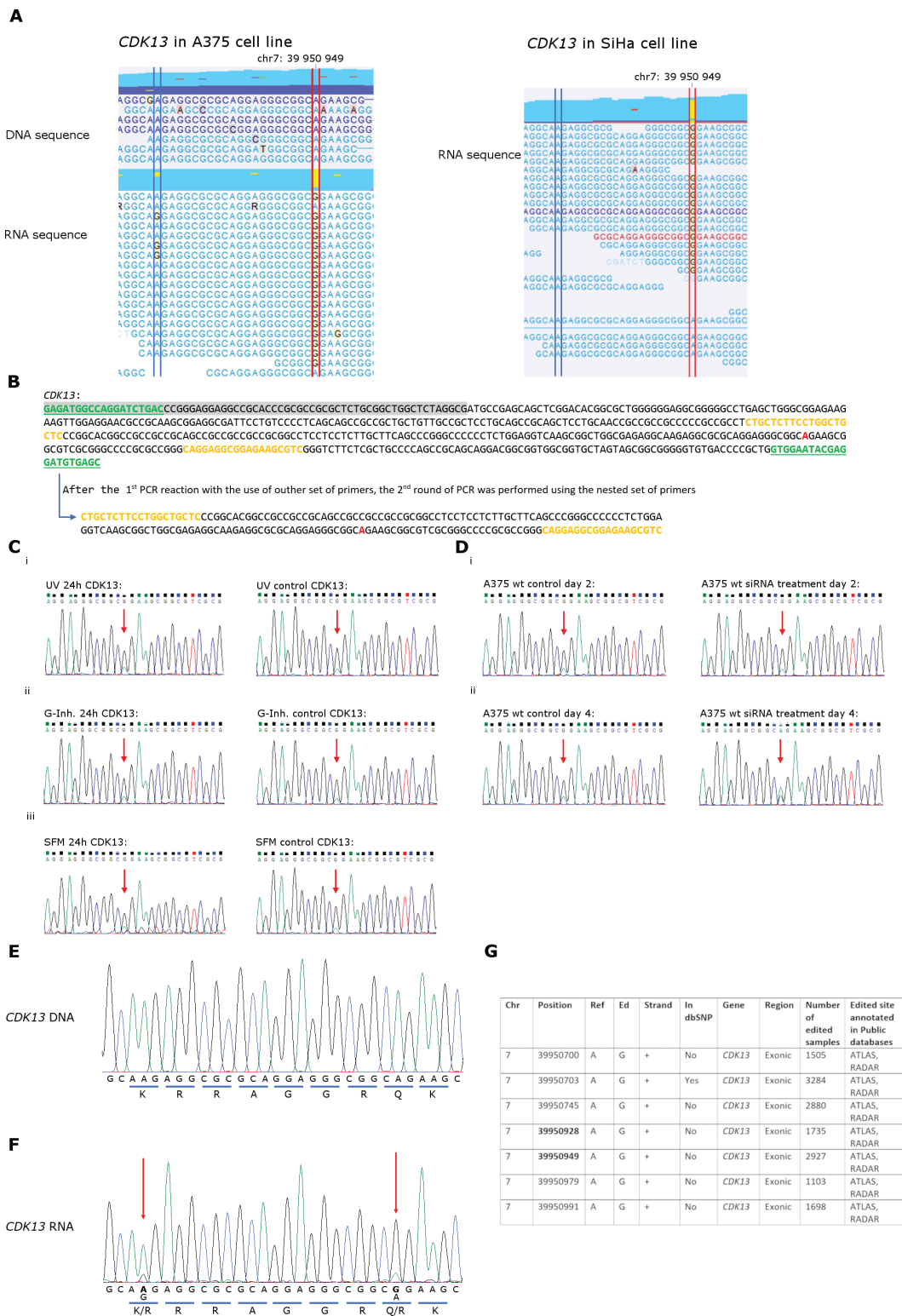
Figure 2. RNA variants that represent potential A-to-I editing events.

(**a**) The *CLC* browser view highlights an A-to-I edit in *CDK13* gene, present in two different cell lines (A375 and SiHa). Only RNA-seq reads are available in the SiHa cell line, thus the possibility of DNA encoded G variant in the sequence cannot be completely excluded. The dominant *CDK13* variant is located on chr7: 39,950,949, p. Q103R, c.308A>G based on Ensembl_v91 and GRCh38. The other A-to-I edit in *CDK13* is indicated by blue lines at on chr7: 39,950,928 and is present only in A375 cells. (**b**) A schematic of the PCR reaction used to amplify the fragment of *CDK13* cDNA with the edited A (marked in red). To overcome problems with GC-reach sequence amplification, nested PCR reaction was optimized first with the use of the outer set of primers (marked in green, forward primer located in 5′UTR). The inner set of primers (marked in yellow) was used to amplify the desired fragment from the *CDK13* cDNA. The 5′UTR is indicated by grey background colour and the A-to-G variant location is marked in red. (**c**) Sequencing results of nested PCR products from the *CDK13* gene using RNA derived from wt-p53 A375 or p53-null A375 cell lines under different cell stress conditions: UV treatment (UV), splicing inhibitor treatment with Isoginkgetin (G-lnh.), and serum starvation (SFM), harvested after 24 hours of treatment or control conditions. All chromatograms show a mixture of edited (G) and non-edited (A) versions of *CDK13* and the ratio between edited and non-edited variants shows no visible quantifiable difference in all samples. The A-to-I edit location is marked by an arrow in the Sanger sequencing reads where the ratio of A (green) to G (black) is visible. (**d**) Sequencing results of nested PCR products from the *CDK13* gene using RNA derived from wt-p53 A375 cells after treatment of cells with an siRNA vs control untreated. siRNA was tested as a 'stress' agent because it is known to induce an interferon response that can impact on

RNA editing. Two days after transfection with siRNA, there is no difference in the ratio between edited and non-edited *CDK13* versions. However, after four days the siRNA treated sample shows non-edited (Adenine, in green) as a predominant variant of the *CDK13* gene. KD, knock-down; A-to-I edit location is marked by an arrow. Sanger sequencing of DNA (**e**) and RNA (**f**) as a validation of NGS data from part A. Red arrows in F indicate the same positions observed in panel E. Noteworthy, the frequency of edit in the first location (chr7: 39,950,928) is lower than the edit in the second location (chr7: 39,950,949), which corresponds to edited reads observed in NGS from (**a**). (**g**) Table presents information of RNA edits deposited in REDIportal within a 290 nt region surrounding analysed RNA edits in *CDK13* gene. Position of edits presented in **a** and **f** panels are in bold. Ref, reference nucleotide; Ed, edited nucleotide; In dbSNP, information if the edit is deposited in dbSNP.

Organism	mRNA sequence number	Source NCBI	Source Ensembl	Comments	Amino acid in isoform
Human (*Homo sapiens*)	XM_011536377.3	V		Predicted* for XP_011534679.1 (isoform X2)	-
	NM_006575.5	V		RefSeq, for NP_006566.2	E
	ENSMUST00000049239.7		V	Present in Ensembl, protein seq: Q9Y4K4.2 (UniProt)	V
Chimpanzee (*Pan troglodytes*)	XM_001155473.5	V		Predicted* for XP_001155473.1	E
	ENSPTRT00000011589.6		V	Present in Ensembl, protein seq: PNI83324.1 (isoform 2)	V
	ENSPTRT00000077988.1		V	Present in Ensembl, protein seq: A0A2I3SF24 (UniProt)	-
Mouse (*Mus musculus*)	XM_006516054.4	V		Predicted* for XP_006516117.1	-
	NM_201519.2	V		RefSeq, for NP_958927.2	E
	ENSMUST00000049239.7		V	Present in Ensembl, protein seq: E9PX30 (UniProt)	A, exists only in Ensembl
Zebrafish (*Danio rerio*)	XM_021480526.1	V		Predicted* for XP_021336201.1	E
	NM_199913.2	V		RefSeq, for NP_956207.1	-

Figure 3. RNA variants in the *MAP4K5* gene.

In the COSMIC database (**a**) the *MAP4K5* codon, marked in yellow, is annotated as a genomic mutation. (**b**) The table presents mRNA sequence number and corresponding amino acid substitution for *MAP4K5* in human, chimpanzee, mouse, and zebrafish. Numbers of sequences deposed in NCBI that are confirmed are indicated in bold. *predicted by automated computational analysis. (**c** and **d**) Two different versions of a reference genome were used to map the variants, (**c**) Ensembl_v74 maps the change to an A-to-T mutation in the RNA or (**d**) Ensembl_v91 maps the RNAseq to a non-canonical 3-nucleotide exon. In (**d**), the TAC-(CTT)-CTA sequencing reads that result in the 3 base exon insertion convert to 5' TAG-(3 base codon)-GTA 3' sequence where the 'AG' forms a classic 3' intron splice acceptor site and the 'GT(U)' forms the classic spice donor site for the subsequent splice to the larger exon in *MAP4K5*. This sequence is the only exact match within this intron, which suggests this is the true exon-intron boundary for this three-base exon. Blue brackets indicate tumour DNA, green brackets indicate tumour RNA. This mRNA sequence deviates from the ensembl reference sequence which encodes a Valine-569 (Fig.4)

editing events localized in this small exonic region and the editing is variable (Fig. 2(g)). The translation of these mRNAs with multiple editing events would, in principle, lead to single, double, triple, and quadruple, etc. mutant CDK13 proteins. As such, functional characterization of these multi-mutant proteins was not evaluated in this study because of the potential variable number of mutant protein products. Nevertheless, our data validate the utility of the software to identify RNA editing events. CDK13 forms an interesting target in the future to validate because the kinase can play a role in splicing regulation by controlling the phosphorylation status and the activity of splicing factors [45]. Furthermore, CDK13 is localized in the nucleus, particularly in speckles which are the storage site for splicing factors [46]. More recently, CDK13 depletion was shown to lead to defects in RNA processing [47]. In recent years, CDK13 has been recognized as a novel oncogene with potent oncogenic activity in various cancer types as it affects cell cycle regulation, proliferation, and chromosome stability functions [48]. Furthermore, pancreatic

Figure 4. Developing Sanger sequencing of two *MAP4K5* isoforms.
(**a**) the sequence highlights the *MAP4K5* cDNA fragment consisting of standard exons 23–26 from Uniprot Q9YK4.2 and XP_011534679.1 containing the exon 23 reference ENSE00003552321 and the exon 24 reference ENSE00003543613. Primers used for amplification are marked in green and underlined. Each new exon starts from the G (in bold). (**b**) The fragment region shown on chromatograms in **d** and **e** with its amino acid translation. (**c**) The scheme below the sequence demonstrates possible splicing variant occurrence. (**d**) The WT-full length *MAP4K5* expression plasmid chromatogram presents the two-exon variant. (**e**) A chromatogram of the RNA variant plasmid expressing the V569E insertion isoform of *MAP4K5* highlights the 3 bp mRNA variant that deviates from the Ensembl reference sequence(as in Fig.3(c and d)). The chromatograms form a reference for the quantitation of mixed mRNA variants derived from biological material in Supplementary Figures 2–4.

disease link associations could be made for upregulated CDK13 by pathway network linkages to p53 [49].

Validation of the non-canonical RNA variant in *MAP4K5* mRNA

The *CLC* RNA variant detection software identifies RNA species that deviate from the reference sequence. This can include not just single nucleotide variants typical of classic RNA editing (A-I or C-U) (Table 1(b)), but also MNVs and indels (Fig. 1(a); Supplementary Tables 1 C-E). Focusing again on protein kinases, using Ensembl RefSeq v74 as a reference sequence, the A-T RNA variant in *MAP4K5* mRNA can be mapped to an exon-intron boundary (Fig. 3 (c)) that is spliced at the downstream intron-exon boundary to create an in frame single amino acid insertion. However an updated version of the RefSeq (Ensembl v91) maps the change to a non-canonical three base exon (Fig. 3(d)). These data

highlight the relative difficulty in mapping very small non-canonical RNA variants to a reference dataset. We think it is more likely that the 'A-T variant' is not mapped correctly to the reference genome and it is more likely that these RNA reads derive from a non-canonical three base exon that deviates from the genomically predicted valine codon at 569 position (Fig. 3(a,b)). In addition, it is unlikely that this three-base non-canonical exon is mapped to the correct intronic position (as reported in Fig. 3(d)) as we can identify several GAA (CTT) motifs across the 2,001 bp intron 23–24.

Regardless of the mechanisms generating the RNA variant in *MAP4K5* mRNA, the genomic annotation of *MAP4K5* at this position is not resolved based on public data in different vertebrate species (Fig. 3(b)). The human MAP4K5 protein found in Ensembl and in NCBI displays three different transcripts with regard to the codon that has the RNA variant (Fig. 3(b) Human). One variant has a reference sequence that codes for the Valine amino acid (GTA). The second transcript has the variant (GAA) which codes for the amino acid
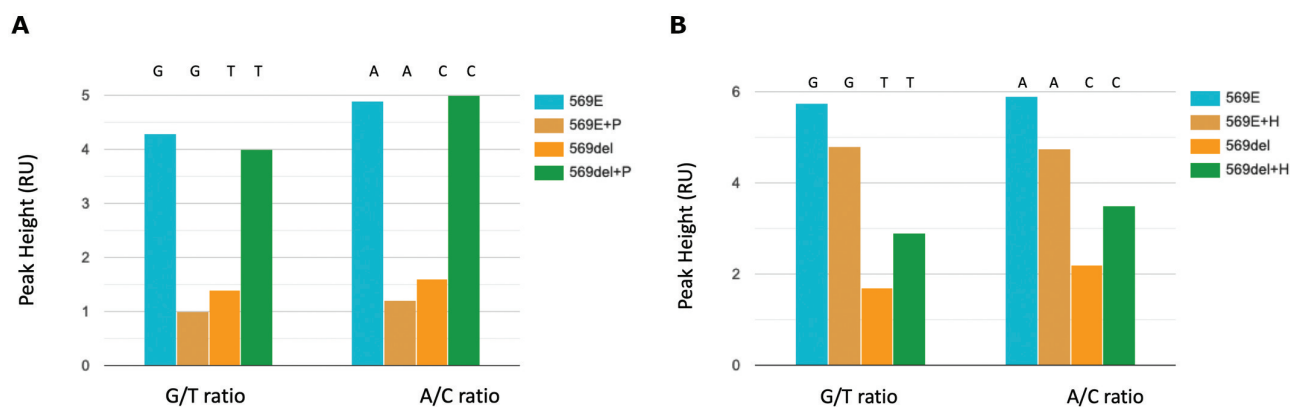
Figure 5. *MAP4K5* Sanger sequencing results quantitation after Pladienolide B and Herboxidiene treatment.
Quantitation of the G-A and T-C sequencing pair at the indicated codon position (* * in Supplementary Figure 3 and Supplementary Figure 4) is quantified in **a** for Pladienolide B treatment and in **b** for Herboxidiene treatment.

Glutamic acid. The third transcript has this codon deleted. Pan species has the same three transcripts to humans (Fig. 3 (b) Chimpanzee). Murine species has the codon GCA, instead of GTA, in one transcript which codes for Alanine (Fig. 3(b) Mouse). The other two transcripts are similar to the second and third transcripts to humans with the variant codon GAA and with the deleted codon, respectively. Danio species have two transcripts only; one with the variant codon GAA, and one with the deleted codon (Fig. 3(b) Zebrafish). Thus, this RNA variant is conserved in many species suggesting selection pressures exist for this specific variant in vertebrates.

We next validated the RNA variant in *MAP4K5* mRNA and whether we could find evidence that it is regulated by splicing mechanisms. First, the reference cDNA sequence predicts a Valine at position 569 based on the Ensembl genomic reference sequence that is encoded by a GTA at the exon23-intron23-24 boundary (Fig. 4(a,b)). If the V569 or E569 allele is derived from the alternative splicing of an exon encoding a Glutamate (E) or Valine (V), then the splicing pattern would be as predicted in Fig. 4(c). We next developed a plasmid expressing MAP4K5 protein including the variant V569E and the 569del allele. Sanger sequencing of the expression plasmid DNAs using primers depicted in Fig. 4(a), gave rise to the SLSGKT amino acid stretch that derives the 569del allele (Fig. 4(d)) or the SLSEGKT that derived the V569E allele (Fig. 4(e)). Having optimized primers for quantifying the RNA variant region, we purified RNA from a range of cells and tissues to determine whether these variants can be detected and quantified. Supplementary Figure 2A reveals that the V569E allele dominates in RNA derived from muscle, adipose, normal human fibroblasts (NHF), or A549 lung cancer cells. In addition, the treatment of A549 lung cancer cells by starvation or over confluence does not alter the expression of the V569E allele (Supplementary Figure 2B). Thus, we see no evidence of the annotated Valine amino acid, although it remains possible that other conditions or cell models might incorporate a different 3 bp exon encoding Valine in place of Glutamate.

We next wanted to determine whether we could drive production of the 569del mRNA species through inhibition of RNA splicing. We used the SF3B1-specific inhibitor Pladienolide B (Fig. 5 and Supplementary Figure 3) [50]. At the 24-h time point, the lower dose of Pladienolide B marginally increased the amount of the 569del allele (Supplementary Figure 3C vs 3B and 3A). However, by 48 hours there was a substantial increase in the 569del allele and reduction in the V569E allele at the lower dose of inhibitor (10 nM; Supplementary Figure 3F vs 3E and 3D). The higher dose of the Pladienolide B inhibitor (100 nM) did not reduce the splicing as affective as the lower dose. Quantification of the Sanger sequencing demonstrated that the ratio of non-spliced:V569E spliced mRNA was 0.32 in non-treated conditions but 4.06 in the presence of Pladienolide B (10 nM; Fig. 5(a)). A different splicing inhibitor (Herboxidiene [50]) that also inhibits SF3B1 was titrated to determine effects on the splicing ratio. This inhibitor also increased the level of the 569del allele which removes the 3 bp Glutamate encoding exon (Supplementary Figure 4(C,F) vs 4A and 4D). Quantification of the Sanger sequencing bases demonstrated that the ratio of non-spliced:V569E spliced mRNA was 0.33 in non-treated conditions but 0.67 in the presence of Herboxidiene B (10 nM; Fig. 5(b)). These data confirm that the RNA variant has emerged due to splicing of the 3 bp exon encoding Glutamate, although further research will be required to map the location of this exon and to determine whether the cell might use the valine encoding 3 bp exon under different physiological conditions.

Having establish that the 569del allele and the V569E allele as the two dominating isoforms detected at the RNA level, we asked whether either of these two variants exhibited any differences in clonogenic activity and/or interaction partners. We would expect some equilibrium shift in the interactome of these two proteins considering the small exon variant is highly conserved within vertebrates. First, the 569del allele and the V569E allele were subcloned into pEXPR-IBA105 containing an SBP tag for affinity purification (Fig. 6(a)). When these two plasmids were transfected into wt-p53 and p53 null melanoma

cell lines, the expression of both could be quantified (Fig. 6 (b)), and the levels of both were equivalently expressed in p53-null cells. These data suggest the variations in the proteins do not dramatically alter their steady state levels. Upon transfection and dilution for limited cell number to measure clonogenicity, there were two notable observations. First, both alleles induced a growth suppression, rather than growth stimulation, suggesting that they function more like tumour suppressors than oncogenes. Second the V569E allele was marginally more active as a growth suppressor, with the V569E suppressing growth by 19% and the 569del suppressing growth by 33% relative to the control vector (Fig. 6(c)). Finally, interactomic immunoprecipitation (IP) identified shared protein–protein interaction sites, but also differences in the quantitative capture of certain protein targets (Fig. 6 (d)). The aggregate of all the significant interacting proteins (above $\log_{0.5}$) that form a network using STRING is shown in Fig. 6(e). Based on these data, the core function of both MAP4K5 isoforms could to be interactions with the protein disulphide isomerase family members including peroxiredoxin (Fig. 6(d), green upper right quadrant). Peroxiredoxin can regulate oxidative stress responses including protein folding [51]. Related to this, the protein RAD23 is also detected in the pull-down experiments and this protein has also been shown to play a role in protein degradation in response to ERAD (endoplasmic reticulum-associated degradation) quality control [52] through interactions with ubiquitin. Based on this core function, and the differential quantitative detection of common interacting proteins, we could speculate that the 569del form of MAP4K5 interacts stronger to the ubiquitin regulator UBA1 and the protein disulphide isomerases – PDIs (PDIA6 and PRDX3) (Fig. 6(d), blue quadrant). By contrast, the V569E isoform shows stronger interactions with secretory or barrier function proteins, such as Fillagrin, XP32, and PRB2 (Fig. 6(d), red quadrants). Together, these data suggest that the 3 bp exon inclusion changes the equilibrium of the kinase towards distinct protein networks and its conservation from fish to mammals suggests this small non-canonical exon plays an important role in MAP4K5 protein function.

Discussion

Cancer genome sequencing has revolutionized our understanding of the genetic basis of cancer, the classes of mutagenic events that drive cancer development, and the identification of genetic drivers [53,54]. However, we do not yet know how this mutated code is translated into an expressed phenotype. Identifying the expressed cancer genome, using RNAseq and protein quantitation methods such as mass spectrometry, provides a more accurate view of the state of the cancer tissue at the time of presentation in the clinic. Although mass spectrometric software including *Proteome Discoverer* or *Maxquant* [55,56] provides a coding-independent user interface that raises the impact of mass spectrometry, the vast majority of next-generation data analysis using DNA variant detectors derived from *Varscan* or *Mutect* requires computational coding [57,58]. An integrated DNA and RNA variant detection software tool (*CLC*) was utilized that is similar in scope to the

Proteome Discoverer software tool used by mass spectrometrists that does not require computational coding [59]. This democratizes DNA and RNA variant detection to the life-science community, from the plant sciences community through to the field of human disease [60–65].

In our prior study, we benchmarked both *CLC* and *Varscan2* as two independent variant detection platforms to define the overlap in their mutation detection and define their dual utility in creating a mutant genomic reference database for optimizing mutant peptide detection using mass spectrometry [66]. Mass spectrometry has also been used previously to identify peptides derived from RNA editing events [21]. We have compared *Varscan* to *CLC* variant detection and have confirmed their similarity at DNA variant identification in a cancer cell model and also confirmed mRNA SNVs at the proteome level through the use of mass spectrometry to identify mutated peptides [66]. Based on the data obtained from the plant science community on RNA-editing using the *CLC* tools, we now have applied this RNA variant detection platform to our human melanoma cell model to determine whether the RNA variant landscape can be defined.

The data output can be exported as the number of synonymous or non-synonymous RNA reads that deviate from or match a reference DNA read set (Supplementary Table 1A and Tables 3 and 4). The RNA sequence deviations from the reference genome can be further annotated into SNVs, MNVs, indels, and replacements (Supplementary Tables 1B-E). In addition, the overall landscape of SNV type can be stratified into A-G/G-A or C-T/T-C mutations (Table 1(b)). In general, the data demonstrate that the ratio of A-G or C-T mutations in RNA is within an order of magnitude of each other. As most RNA editing software focus on a type of RNA editing (such as A-I), it is not clear whether the C-U or A-I RNA editing landscape defined by *CLC* is representative of other platforms or other cell models. For example, even in studies where APOBEC dependent RNA editing is amplified [67,68], such studies do not usually compare the ratio of ADAR to APOBEC events.

In our current study, we first focused on validating one of the A-I RNA-editing events we detected, that has been reported previously. *CDK13* mRNA over-editing has been reported in liver cancer [40] using RNAEditor [69]. This latter study identified two RNA editing events that could give rise to Q103R and K96R mutation in *CDK13*. We also can observe both RNA A-I editing events, with the G sequencing reads detected at codon 103 and 96 positions (in blue and red lines, Fig. 2(a i)) although the Q103R variant predominates (Fig. 2(a)). The SiHa cell model only identified RNA editing changes at the 103 codon position (Fig. 2(a)). It is noteworthy that the other *CDK* orthologues-*CDK10, CDK11*, and *CDK12* which share similar G-C rich regions as *CDK13* also display a degree of A-I editing in their transcripts (Table 2, *CDK10* (Gln46Arg); *CDK11* (Cys23Arg); and *CDK12* (Gly1425Arg)). It will be interesting in the future to expand on the editing regulation of these *CDK* orthologues, determine whether mutant proteins are produced as a result of the edit, and what might be the change in function of the edited genes.

In addition to classic A-I or C-U RNA edits, the *CLC* software also defines RNA variations in a cDNA sequence
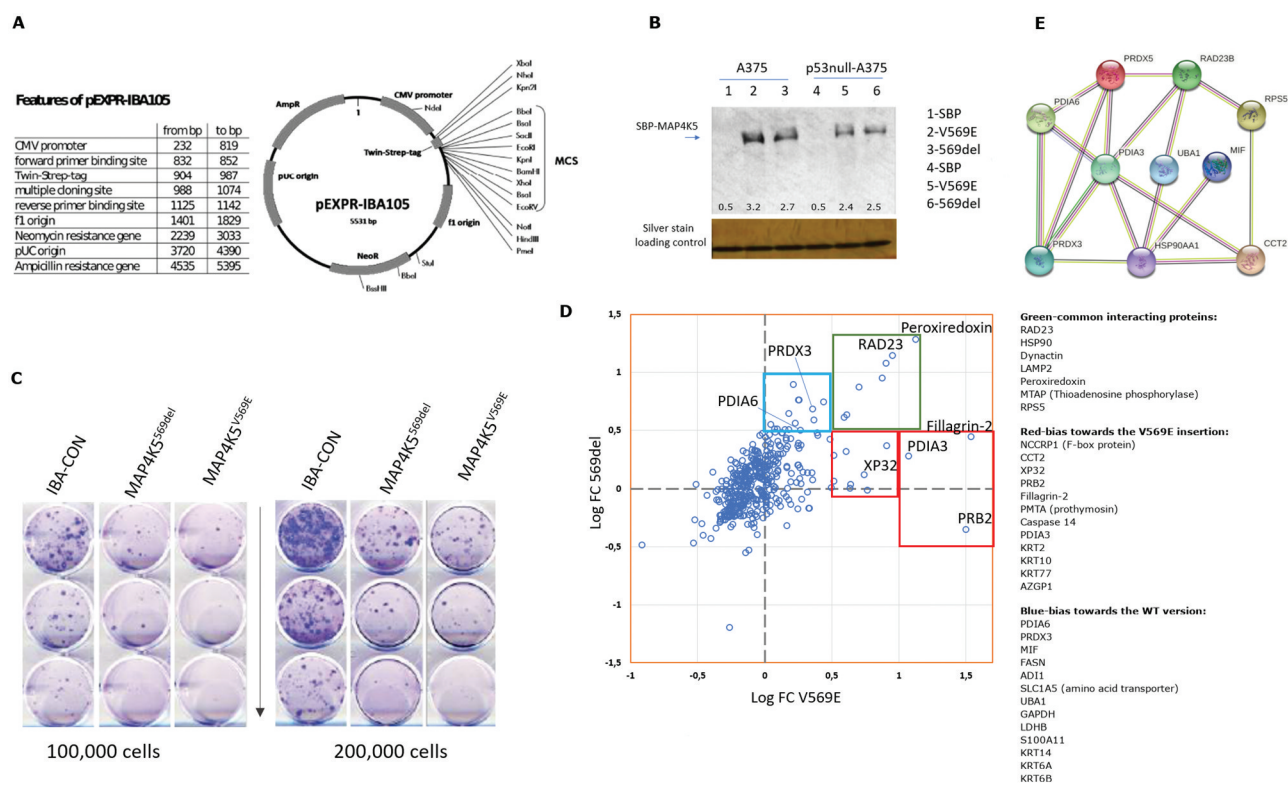
Figure 6. Biochemical evaluation of *MAP4K5* variants.

(**a**) A description and a map of the expression vector pEXPR-IBA105 containing both wild-type (V569del) or the V569E insertion variant of *MAP4K5*. (**b**) A blot presenting SBP-tagged MAP4K5^{569del} and MAP4K5^{V569E} in wt-p53 A375 or p53-null A375 cells. After transfection, the material was then immunoblotted with anti-SBP antibody and in parallel the loading was analysed using silver staining of the transfected lysates. (**c**) Clonogenic assay with the use of the empty IBA-CON vector (control), SBP-tagged MAP4K5^{569del}, and MAP4K5^{V569E}. (**d**) Immunoprecipitation results of A375 p53 WT transfected with MAP4K5^{V569E} (V569E, presented on x-axis) and MAP4K5^{569del} (569del, shown on y-axis) vectors, presented as fold change in log scale (Log(FC)). Proteins that show similar high co-immunoprecipitation in both MAP4K5 versions are indicated in green quadrant, proteins commonly co-immunoprecipitated but enriched with MAP4K5^{V569E} but not as high with MAP4K5^{569del} are marked in red quadrant, and proteins commonly co-immunoprecipitated but higher with MAP4K5^{569del} relative to MAP4K5^{V569E} are marked in blue quadrant. (**e**) STING pathway showing a network of common proteins interacting with MAP4K5^{V569E} and MAP4K5^{569del} variants.

that does not match a reference genome. Genes like *MAP4K5* or *DIAPH2* (Supplementary Table 1C) have in frame small exons (3–6 bp) that represent 'unmapped' or non-canonical exons that are difficult to annotate with large intronic regions. We focus here on validating one representative class with a small in-frame exon, *MAP4K5*. The genomic sequence of *MAP4K5* predicts a GTA sequence at exon23-intron23-24 boundary that encodes a Valine at codon 569 (Figs. 3 and 4). However, the RNA sequencing produces data that is defined as a variant by the *CLC* software (Fig. 3) that replaces a GTA with GAA leading to an Glutamate at codon 569. However, the genome sequence cannot accommodate this unless an A-T edit is annotated (Fig. 3(c)) or a downstream GAA exon (CTT reverse strand) is annotated into the intron (Fig. 3(d)). The A-T deviation can be altered using splicing inhibitors that reduces the GAA sequence (Fig. 5 and Supplementary Figures 3 and 4) suggesting that the deviation comes from a 3 bp exon in the intron. However, as a 1 bp exon has been reported previously [70], we cannot be confident whether the GAA exon arises from a 3 bp exon within intron 23–24, or a 1 bp exon that fuses to the terminal end of exon 23 and deletion of T nucleotide on the beginning of exon

24. Nevertheless, so-called microexons have been observed previously with as little as 3 bp exons being detected [71]. In conclusion, the integrated DNA and RNA variant detection software described in our study can open the door to more routine analysis of these splicing phenomenon by the life-science community and support future analysis of RNA variant detection in cancer tissue.

Methods

Cells and reagents

All chemicals were obtained from Sigma Aldrich unless otherwise indicated. A375 cells were reported previously [72]. The p53-specific gRNA sequence was 5′-CTGAGCAGCGCTCATGGTGGNGG-3′, and was used to develop the isogenic p53-null cell line as reported previously [73]. A549 cells were described previously [74]. The cell line from ATCC was cultured in DMEM (Gibco) + 10% FBS (Gibco) medium. Cells were split every 2 days, 0.05% Trypsin-EDTA (Gibco) was used to detach cells. The GI50 for cell growth inhibition for the splicing inhibitor Herboxidiene

(Focus Biomolecules) [50,75,76]. The anti-SBP tagged antibody was form IBA (Mouse, Monoclonal). The HRP-conjugated Anti-mouse (was PO260 from Dako) and the HRP-conjugated Anti-rabbit (was PO217 from Dako).

DNA and RNA sequencing

Exome Sequencing of DNA derived from A375 cells was performed using Agilent V5+ UTR Exome Capture Kit (75Mb) and 100 bp paired-end reads were acquired using a coverage of 100x (performed by Otogenetics, USA). The paired fastq files from the A375 cell line (available upon request) from exome sequencing were imported into the *CLC Biomedical Genomics Workbench*. Adaptor sequences and bases with low quality were trimmed, DNA sequencing reads were mapped to the human reference genome hg19. Adaptor sequences and bases with low quality were trimmed and reads were mapped to human genome 19 (hg19). Variants were detected in the exome data with the *CLC* Probabilistic Variant Caller using the following parameters: Minimum coverage (number of reads) = 5; Minimum frequency = 5%; Minimum number of variants = 2; Variants in normal germline DNA = 0, and the coverage in the germline DNA should be at least 5 reads at the variant site. Sequencing of RNA derived from both wt-p53 and p53-null A375 cell panel, untreated and treated with IFNγ (1 ng/ml for 24 hours), to generate biological replicates based on the common genomic DNA reference file from the parental A375 cell line, was performed using total RNA, depleted of ribosomal RNA, followed by random priming to generate cDNA. From this template paired-end Illumina HiSeq2500 was used to generate approximately 20 million reads. Paired fastq files (available upon request) from RNAseq reads were imported into the *CLC Biomedical Genomics Workbench*. The RNA sequencing reads were mapped to the human reference genome hg19. Paired de-multiplexed fastq files from RNAseq libraries were trimmed for stretches of adapter sequences, joined into a single read followed by quality trimming using commands from the CLC Assembly Cell. The fastq files were then imported into the *CLC Biomedical Genomics Workbench* (version 2.5). Sequences were mapped to the A375 cancer genome sequence where at least 2 mutant RNA reads were identified that do not match the reference genomic DNA.

MAP4K5 immunoprecipitation and SWATH-MS

A plasmid containing the MAP4K^{V569E} gene was acquired from Addgene (Addgene plasmid # 23,611) and the gene was cloned in pEXPR-IBA105 vector containing Streptavidin Binding Peptide (SBP). The MAP4K^{V569E} expression plasmid was subsequently mutagenized using the DpnI method [77] to create a deletion of 3 base pairs (at codon 569) to obtain the MAP4K^{569del} form (Fig.6 (a)). The primers included; MAP4K5 cloning (F, with EcoRI restriction site): 5′-GTCCCGAATTCGATGGAGGCCCCGCTG -3′; MAP4K5 cloning (R, with BamHI restriction site): 5′-CCCGGGGATCCCTTAGTAACTATTTTCATGTCCAGCCA-AGAT3′; MAP4K5 mutagenesis (F): 5′-ATTATCAGGAAAAACCTTTCAGC-3′; and MAP4K5

mutagenesis (R): 5′-AGCTGAAAGGTTTTTCCTGATAAT-3′. The MAP4K5 isoforms (Fig.6(a-b)) were transfected into A375 cells (as in Fig.6(b)) and immunoprecipitation (IP) was carried out with the MAP4K5-expression vectors and the SBP empty vector (pEXPR-IBA105). Cells were transfected at about 70–80% of confluency using Attractene, harvested the next day, and lysed with Triton lysis buffer (100 *mM* KCl, 20 *mM* HEPES pH7.5, 1 *mM* EDTA, 1 *mM* EGTA, 0.5 *mM* Na3VO4, 10% glycerol, 0.5X Protease Inhibitor Mix, 10 *mM* NaF, 0.1% Triton x-100). 30 µl of streptavidin agarose conjugate beads (Millipore) was washed three times with 500 µl of PBS and lysate was added to the beads and incubated on a rotor wheel for 2 hours at room temperature. After the incubation, the sample was washed one time with 500 µl of Triton lysis buffer (without Triton x-100) and two times with 500 µl of PBS. Finally, the sample was eluted in 120 µl of elution buffer (8 *M* urea, 2 *mM* DTT and 20 *mM* HEPES pH 8) incubating at 85°C for 5 *min*.

The lysates was processed by the method of FASP [78] to obtain tryptic peptides. FASP-processed tryptic peptides from the streptavidin bead pull-down were separated on an Eksigent Ekspert nanoLC 400 (SCIEX, California, USA) online connected to a TripleTOF 5600+ (SCIEX, Toronto, Canada) mass spectrometer. Data acquisition was performed in technical triplicates. A cartridge trap column (300 µm i.d. × 5 mm) packed with a C18 PepMap100 sorbent with a 5 µm particle size (Thermo Fisher Scientific, Waltham, MA, USA) was used to concentrate and wash peptides. Peptides were washed in 0.05% trifluoroacetic acid in 5% acetonitrile and 95% water for 10 minutes. Following, peptides were separated using a gradient of acetonitrile/water (300nL/minute) on an analytical capillary emitter column PicoFrit® (75 µm × 210 mm (New Objective, Massachusetts, USA)) self-packed with ProntoSIL 120-3-C18 AQ sorbent with 3 µm particles (Bischoff, Leonberg, Germany). Analytical gradient was mixed from Mobile phase A composed from 0.1% (v/v) formic acid in water, and mobile phase B composed of 0.1% (v/v) formic acid in acetonitrile. Gradient elution started at 5% mobile phase B for the first 30 minutes and then the proportion of mobile phase B increased linearly up to 40%B for the following 120 minutes. Output from the separation column was directly coupled to an ion source (nano-electrospray).

The spectral library sample was prepared by pooling equal volume (10 µl) of all samples followed by data-dependent shotgun measurement in positive mode (IDA). Precursor range in MS was set from m/z 400 up to m/z 1250 while MS/MS spectra were acquired from m/z 200 up to m/z 1600. DIA method acquired and fragmented 20 the most intensive precursor ions in each cycle. Cycle time was 2.3 seconds. Once measured precursors were excluded for 12 seconds. Protein identification was performed using Protein Pilot 4.5 (SCIEX, Toronto, Canada) search engine. Acquired MS and MS/MS spectra were searched against Uniprot+Swissprot database (02. 2016, 69,987 entries) restricted to *Homo sapiens* taxonomy. Alkylation on cysteine using iodoacetamide as a fixed modification and digestion using trypsin was specified in the search engine. A decoy database was generated to perform FDR analysis. A spectral library was generated in Peakview 1.2.0.3 (SCIEX, Toronto, Canada) from DIA search files where

only proteins with FDR below 1% were imported into the spectral library. SWATH measurements were operated in positive high sensitivity mode. Precursor ions spanning from m/z 400 up to m/z 1200 were measured in windowed manner. Precursor mass range was divided into 67 precursor windows with 12 Da width and 1 Da overlap. Accumulation time of 50 ms was set per each SWATH precursor window resulting in 3.0 seconds cycle time. Product ions mass range was set from m/z 400 up to m/z 1600. SWATH data extraction was performed in Peakview 1.2.0.3 (SCIEX, Toronto, Canada) with the spectral library. The retention time window for extraction was manually set to 10 minutes. Up to 4 peptides and 6 product ions per each peptide were used to quantitate each protein. Only non-modified high confidence peptides (peptide confidence>99%) were used for quantitation. Protein summed peak areas were determined from the sum of corresponding transition peak areas. Normalization was performed using the total area sums option in MarkerView 1.2.1.1 (SCIEX, Toronto, Canada). Extracted quantitative data from three technical replicates were statistically evaluated in MarkerView 1.2.1.1 (AB-SCIEX, Canada). Pairwise T-test was performed to determine protein fold changes and P values of fold change for all proteins listed in the spectral library.

Pladienolide B and Herboxidiene treatment

A549 cells were treated with Pladienolide B (10 or 100 nM) or Herboxidiene (5 or 10 nM) dissolved in full-medium for 24 and 48 hours (Pladienolide B) or for 48 and 72 hours (Herboxidiene). Control samples contained DMSO alone. After treatment, cells were harvested and total RNA was extracted. RNA was isolated from cells using Universal RNA Purification kit (EURex, cat no. E3598), according to the manufacturer's protocol. The RNA concentration was measured with the use of NanoReady (Life Real) device. Reverse transcription was performed in 20 μl reaction volume, with the use of High Capacity cDNA Reverse Transcription Kit (REF: 4,368,814, Thermo Fisher Scientific), according to the manufacturer's protocol. 500ng of RNA was used for this reaction.

PCR amplification and purification

Amplification of *MAP4K5* fragment was performed using Phusion High-Fidelity DNA Polymerase kit (Thermo Fisher Scientific) using 100 ng cDNA as a template and subsequent primer sequences: M4K5e23F: TCCACGGAAGTGTACTTGGC; M4K5e26R: TCCAGACTGTAAAGCTCCACA. The thermal cycler program for *MAP4K5* gene amplification was as follows: 1. Denaturation: 95°C, 3 min; 2. Denaturation: 98°C, 20 sec; 3. Annealing: 60°C, 15 sec; 4. Elongation: 72°C, 20 sec; Steps 2–4 were repeated 29 times.; 5. Elongation: 72°C, 2 min; and 6. Hold: 4°C, Inf.

Amplification of *CDK13* fragment was performed using Phusion MM in HF buffer kit (Thermo Fisher Scientific) using 100 ng cDNA as a template and primer sequences for the 1st PCR reaction: CDK13 F3.3: GAGATGGCCAGGATCTGAC; CDK13 R1: GTGGAATACGAGGATGTGAGC. As a template for the 2nd

PCR reaction 100 ng of purified (with the use of QIAquick PCR Purification Kit (QIAGEN)) product from the 1st PCR was used and the primer sequences were as follows: CDK13 F1: CTGCTCTTCCTGGCTGCTC and CDK13 R2: CAGGAGGCGGAGAAGCGTC. The thermal cycler program for both PCR reactions was as follows: for the 1st PCR: 1. Denaturation: 98°C, 2 min; 2. Denaturation: 98°C, 45 sec; 3. Annealing: 63°C, 30 sec; 4. Elongation: 72°C, 30 sec; Steps 2–4 were repeated 10 times.; 5. Elongation: 72°C, 10 min; and 6. Hold: 4°C, Inf.; for 2nd PCR: 1. Denaturation: 98°C, 2 min; 2. Denaturation: 98°C, 45 sec; 3. Annealing: 61°C, 30 sec; 4. Elongation: 72°C, 30 sec; Steps 2–4 were repeated 25 times.; 5. Elongation: 72°C, 10 min; and 6. Hold: 4°C, Inf.

The PCR products were visualized on 1.5% agarose gel with the use of 1kb Gene Ruler (Thermo Fisher Scientific). The purification of PCR products was performed using NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel) and the concentration of purified PCR products was measured on Nanodrop or NanoReady device. Purified PCR products were sequenced to the Eurofins company. The results (chromatograms) were visualized in Chromas v2.6.6.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

ORCID

Małgorzata Kurkowiak http://orcid.org/0000-0002-4968-7136
Jakub Faktor http://orcid.org/0000-0002-8232-9787
Natalia Marek-Trzonkowska http://orcid.org/0000-0003-1201-295X
Ted R. Hupp http://orcid.org/0000-0003-0391-0352

References

[1] Gillette MA, Satpathy S, Cao S, et al. Clinical proteomic tumor analysis, proteogenomic characterization reveals therapeutic vulnerabilities in lung Adenocarcinoma. Cell. 2020;182:200–225 e235.

[2] McDermott JE, Arshad OA, Petyuk VA. et al. Clinical tumor analysis, proteogenomic characterization of ovarian HGSC implicates mitotic kinases, replication stress in observed chromosomal instability. 2020;1. Cell Rep Med

[3] Dou Y, Kawaler EA, Zhou DC, et al. Clinical proteomic tumor analysis, proteogenomic characterization of endometrial carcinoma. Cell. 2020;180(4):729–748 e726. .

[4] Binetti M, Lauro A, Vaccari S, et al. Proteogenomic biomarkers in colorectal cancers: clinical applications. Expert Rev Proteomics. 2020;17(5):355–363.

[5] Clark DJ, Dhanasekaran SM, Petralia F, et al. Clinical proteomic tumor analysis, integrated proteogenomic characterization of clear cell renal cell carcinoma. Cell. 2020;180(1):207.

[6] Calinawan AP, Song X, Ji J, et al. ProTrack: an interactive Multi-Omics data browser for proteogenomic studies. Proteomics. 2020;20(21–22):e1900359.

[7] Alexandrov LB, Jones PH, Wedge DC, et al. Clock-like mutational processes in human somatic cells. Nat Genet. 2015;47 (12):1402–1407.

[8] Nik-Zainal S, Kucab JE, Morganella S, et al. The genome as a record of environmental exposure. Mutagenesis. 2015;30(6):763–770.

[9] An B, Pan T, Hu J, et al. The discovery of a potent and selective third-generation EGFR kinase inhibitor as a therapy for EGFR L858R/T790M double mutant non-small cell lung cancer. Eur J Med Chem. 2019;183:111709.

[10] Piper-Vallillo AJ, Sequist LV, Piotrowska Z. Emerging treatment paradigms for EGFR-Mutant lung cancers progressing on Osimertinib: a review. J Clin Oncol. 2020;JCO1903123. doi: 10.1200/JCO.19.03123. Online ahead of print.

[11] Yadav M, Jhunjhunwala S, Phung QT, et al. Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. Nature. 2014;515(7528):572–576.

[12] Schumacher T, Bunse L, Pusch S, et al. A vaccine targeting mutant IDH1 induces antitumour immunity. Nature. 2014;512 (7514):324–327.

[13] Carreno BM, Magrini V, Becker-Hapak M, et al. Cancer immunotherapy. A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific T cells. Science. 2015;348(6236):803–808.

[14] Schlake T, Thess A, Fotin-Mleczek M, et al. Developing mRNA-vaccine technologies. RNA Biol. 2012;9(11):1319–1330.

[15] Colloca S, Barnes E, Folgori A, et al. Vaccine vectors derived from a large collection of simian adenoviruses induce potent cellular immunity across multiple species. Sci Transl Med. 2012;4 (115):115ra112. .

[16] Leoni G, D'Alise AM, Cotugno G, et al. A genetic vaccine encoding shared cancer neoantigens to treat tumors with microsatellite instability. Cancer Res. 2020;80(18):3972–3982. .

[17] Zhang B, Wang J, Wang X, et al. Proteogenomic characterization of human colon and rectal cancer. Nature. 2014;513 (7518):382–387. .

[18] Noorani A, Li X, Goddard M, et al. Genomic evidence supports a clonal diaspora model for metastases of esophageal adenocarcinoma. Nat Genet. 2020;52(1):74–83. .

[19] Ruggles KV, Krug K, Wang X, et al. Methods, tools and current perspectives in proteogenomics. Mol Cell Proteomics. 2017;16 (6):959–981.

[20] Apcher S, Daskalogianni C, Fahraeus R. Pioneer translation products as an alternative source for MHC-I antigenic peptides. Mol Immunol. 2015;68(2):68–71.

[21] Zhang M, Fritsche J, Roszik J, et al. RNA editing derived epitopes function as cancer antigens to elicit immune responses. Nat Commun. 2018;9(1):3919. .

[22] Chen X, Nagai Y, Zhu Z, et al. A spliced form of CD44 expresses the unique glycan that is recognized by the prostate cancer specific antibody F77. Oncotarget. 2018;9(3):3631–3640.

[23] Krasnov GS, Dmitriev AA, Kudryavtseva AV, et al. PPLine: an automated pipeline for SNP, SAP, and splice variant detection in the context of proteogenomics. J Proteome Res. 2015;14 (9):3729–3737.

[24] Jagtap P, Goslinga J, Kooren JA, et al. A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. Proteomics. 2013;13(8):1352–1357.

[25] Choi S, Ju S, Lee J, et al. Proteogenomic Approach to UTR Peptide Identification. J Proteome Res. 2020;19(1):212–220.

[26] Faktor J, Grasso G, Vojtesek V, et al. The effects of p53 gene deletion on mutant proteome expression. BBA General Subjects. 2020;1864(12):129722. doi: 10.1016/j.bbagen.2020.129722. Online ahead of print.

[27] Knutson SD, Heemstra JM. EndoVIPER-seq for improved detection of A-to-I Editing sites in cellular RNA. Curr Protoc Chem Biol. 2020;12(2):e82.

[28] Picardi E, D'Erchia AM, Lo Giudice C, et al. REDIportal: a comprehensive database of A-to-I RNA editing events in humans. Nucleic Acids Res. 2017;45(D1):D750–D757.

[29] Schaffer AA, Kopel E, Hendel A, et al. The cell line A-to-I RNA editing catalogue. Nucleic Acids Res. 2020;48(11):5849–5858.

[30] Lo Giudice C, Tangaro MA, Pesole G, et al. Investigating RNA editing in deep transcriptome datasets with REDItools and REDIportal. Nat Protoc. 2020;15(3):1098–1131.

[31] Lo Giudice C, Pesole G, Picardi E. REDIdb 3.0: a comprehensive collection of RNA editing events in plant organellar genomes. Front Plant Sci. 2018;9:482.

[32] Ramaswami G, Li JB. RADAR: a rigorously annotated database of A-to-I RNA editing. Nucleic Acids Res. 2014;42(D1):D109–113.

[33] Brenner WG, Mader M, Muller NA, et al. High level of conservation of mitochondrial RNA editing sites among four populus species. G3 (Bethesda). 2019;9(3):709–717.

[34] Zhou ZY, Hu Y, Li A, et al. Genome wide analyses uncover allele-specific RNA editing in human and mouse. Nucleic Acids Res. 2018;46(17):8888–8897.

[35] Forbes C. Microcomputer programs for mutation studies using the Fisher exact test or the binomial approximation. Mutat Res. 1984;141(3–4):205–210.

[36] Niavarani A, Currie E, Reyal Y, et al. APOBEC3A is implicated in a novel class of G-to-A mRNA editing in WT1 transcripts. PLoS One. 2015;10(3):e0120089.

[37] Castandet B, Araya A. RNA editing in plant organelles. Why make it easy? Biochemistry (Mosc). 2011;76(8):924–931.

[38] Tassinari V, Cesarini V, Silvestris DA, et al. MicroRNA editing detection and function: a combined in silico and experimental approach for the identification and validation of putative oncogenic targets. Methods Mol Biol. 2021;2181:253–267.

[39] Yuting K, Ding D, Iizasa H. Adenosine-to-Inosine RNA Editing Enzyme ADAR and microRNAs. Methods Mol Biol. 2021;2181:83–95.

[40] Dong X, Chen G, Cai Z, et al. CDK13 RNA Over-Editing mediated by ADAR1 associates with poor prognosis of hepatocellular carcinoma patients. Cell Physiol Biochem. 2018;47(6):2602–2612.

[41] Li W, Lynch M. Universally high transcript error rates in bacteria. Elife. 2020;9:e54898.

[42] Ma X, Shao Y, Tian L, et al. Analysis of error profiles in deep next-generation sequencing data. Genome Biol. 2019;20(1):50.

[43] Chepelev I. Detection of RNA editing events in human cells using high-throughput sequencing. Methods Mol Biol. 2012;815:91–102.

[44] Maas S, Godfried Sie CP, Stoev I, et al. Genome-wide evaluation and discovery of vertebrate A-to-I RNA editing sites. Biochem Biophys Res Commun. 2011;412(3):407–412.

[45] Even Y, Escande ML, Fayet C, et al. CDK13, a Kinase involved in Pre-mRNA splicing, is a component of the perinucleolar compartment. PLoS One. 2016;11(2):e0149184.

[46] Even Y, Durieux S, Escande ML, et al. CDC2L5, a Cdk-like kinase with RS domain, interacts with the ASF/SF2-associated protein p32 and affects splicing in vivo. J Cell Biochem. 2006;99(3):890–904.

[47] Liang K, Gao X, Gilmore JM, et al. Characterization of human cyclin-dependent kinase 12 (CDK12) and CDK13 complexes in C-terminal domain phosphorylation, gene transcription, and RNA processing. Mol Cell Biol. 2015;35(6):928–938.

[48] Kim HE, Kim DG, Lee KJ, et al. Frequent amplification of CENPF, GMNN and CDK13 genes in hepatocellular carcinomas. PLoS One. 2012;7(8):e43223.

[49] Ansari D, Andersson R, Bauden MP, et al. Protein deep sequencing applied to biobank samples from patients with pancreatic cancer. J Cancer Res Clin Oncol. 2015;141(2):369–380.

[50] Cretu C, Agrawal AA, Cook A, et al. Structural basis of splicing modulation by antitumor macrolide compounds. Mol Cell. 2018;70(2):265–273 e268.

[51] Sharapov MG, Gudkov SV. Peroxiredoxin 1 - Multifunctional antioxidant enzyme, protects from oxidative damages and increases the survival rate of mice exposed to total body irradiation. Arch Biochem Biophys. 2021;697:108671.

[52] Hanzelmann P, Stingele J, Hofmann K, et al. The yeast E4 ubiquitin ligase Ufd2 interacts with the ubiquitin-like domains of Rad23 and Dsk2 via a novel and distinct ubiquitin-like binding domain. J Biol Chem. 2010;285(26):20390–20398.

[53] Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646–674.

[54] Alexandrov LB, Kim J, Haradhvala NJ, et al., . Consortium, The repertoire of mutational signatures in human cancer. Nature. 2020;578(7793):94–101.

[55] Tyanova S, Temu T, Cox J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. Nat Protoc. 2016;11(12):2301–2319.

[56] Gorshkov V, Verano-Braga T, Kjeldsen F. SuperQuant: a data processing approach to increase quantitative proteome coverage. Anal Chem. 2015;87(12):6319–6327.

[57] Do Valle IF, Giampieri E, Simonetti G, et al. Optimized pipeline of MuTect and GATK tools to improve the detection of somatic single nucleotide polymorphisms in whole-exome sequencing data. BMC Bioinformatics. 2016;17(S12):341.

[58] Cai L, Yuan W, Zhang Z, et al. In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data. Sci Rep. 2016;6(1):36540.

[59] Hedegaard J, Thorsen K, Lund MK, et al. Next-generation sequencing of RNA and DNA isolated from paired fresh-frozen and formalin-fixed paraffin-embedded samples of human cancer and normal tissue. PLoS One. 2014;9(5):e98187.

[60] Wang W, Ma L, Becher H, et al. Astonishing 35S rDNA diversity in the gymnosperm species Cycas revoluta Thunb. Chromosoma. 2016;125(4):683–699.

[61] Sujeeth N, Mehterov N, Gupta S, et al. A novel seed plants gene regulates oxidative stress tolerance in Arabidopsis thaliana. Cell Mol Life Sci. 2020;77(4):705–718.

[62] Roberts SA, Sterling J, Thompson C, et al. Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. Mol Cell. 2012;46(4):424–435.

[63] Yeh YS, Chang YT, Ma CJ, et al. First-decade patient with colorectal cancer carrying both germline and somatic mutations in APC gene. BMC Cancer. 2017;17(1):849.

[64] Chen TC, Liu YC, Wang X, et al. Whole plastid transcriptomes reveal abundant RNA editing sites and differential editing status in Phalaenopsis aphrodite subsp. formosana. Bot Stud. 2017;58(1):38.

[65] Villarreal AJ, Turmel M, Bourgouin-Couture M, et al. Genome-wide organellar analyses from the hornwort Leiosporoceros dussii show low frequency of RNA editing. PLoS One. 2018;13(8):e0200491.

[66] Faktor J, Grasso G, Zavadil Kokas F, et al. The effects of p53 gene inactivation on mutant proteome expression in a human melanoma cell model. Biochim Biophys Acta Gen Subj. 2020;1864(12):129722.

[67] Valdmanis PN, Roy-Chaudhuri B, Kim HK, et al. Upregulation of the microRNA cluster at the Dlk1-Dio3 locus in lung adenocarcinoma. Oncogene. 2015;34(1):94–103.

[68] Yamanaka S, Balestra ME, Ferrell LD, et al. Apolipoprotein B mRNA-editing protein induces hepatocellular carcinoma and dysplasia in transgenic animals. Proc Natl Acad Sci U S A. 1995;92(18):8483–8487.

[69] John D, Weirick T, Dimmeler S, et al. RNAEditor: easy detection of RNA editing events and the introduction of editing islands. Brief Bioinform. 2017;18(6):993–1001.

[70] Guo L, Liu CM. A single-nucleotide exon found in Arabidopsis. Sci Rep. 2015;5(1):18087.

[71] Irimia M, Weatheritt RJ, Ellis JD, et al. A highly conserved program of neuronal microexons is misregulated in autistic brains. Cell. 2014;159(7):1511–1523.

[72] Blaydes JP, Luciani MG, Pospisilova S, et al. Stoichiometric phosphorylation of human p53 at Ser315 stimulates p53-dependent transcription. J Biol Chem. 2001;276(7):4699–4708.

[73] Li R, Zatloukalova P, Muller P, et al. The MDM2 ligand Nutlin-3 differentially alters expression of the immune blockade receptors PD-L1 and CD276. Cell Mol Biol Lett. 2020;25(1):41.

[74] Gray TA, Alsamman K, Murray E, et al. Engineering a synthetic cell panel to identify signalling components reprogrammed by the cell growth regulator anterior gradient-2. Mol Biosyst. 2014;10(6):1409–1425.

[75] Teng T, Tsai JH, Puyang X, et al. Splicing modulators act at the branch point adenosine binding pocket defined by the PHF5A-SF3b complex. Nat Commun. 2017;8(1):15522.

[76] Siebring-van Olst E, Blijlevens M, De Menezes RX, et al. A genome-wide siRNA screen for regulators of tumor suppressor p53 activity in human non-small cell lung cancer cells identifies components of the RNA splicing machinery as targets for anticancer treatment. Mol Oncol. 2017;11(5):534–551.

[77] Forloni M, Liu AY, Wajapeyee N. Multisite-Directed Mutagenesis. Cold Spring Harb Protoc. 2019;2019(12). DOI:10.1101/pdb.prot097816

[78] Wisniewski JR, Zougman A, Mann M. Combination of FASP and StageTip-based fractionation allows in-depth analysis of the hippocampal membrane proteome. J Proteome Res. 2009;8(12):5674–5678.

[79] Zhu Y, Luo H, Zhang X, et al. Abundant and selective RNA-Editing events in the medicinal mushroom Ganoderma lucidum. Genetics. 2014;196(4):1047–1057.