Gene fusion between CDKN1A and RAB44 caused by exon skipping like mechanism due to disruption of a splice site

Han Sun[1*], Jingyan Wu[1*], Chenchen Zhu[3], Raeka Aiyar[2], Petra Jakob[3], William Mueller[3], Wu Wei[1, 2§] & Lars M. Steinmetz[1, 2, 3§]

[1]Department of Genetics, School of Medicine, Stanford University, Stanford, CA 94305, USA

[2]Stanford Genome Technology Center, Stanford University, Palo Alto, CA 94304, USA

[3]European Molecular Biology Laboratory (EMBL), Genome Biology Unit, 69117 Heidelberg, Germany

[*] These authors contributed equally to this work

[§]Correspondence should be addressed to Dr. Steinmetz (larsms@stanford.edu) or Dr. Wei (wuwei5@stanford.edu)

Abstract

We anecdotally observed a novel exon skipping event, which was due to the disruption of an exon splicing enhancer far from the exon intron boundary, when doing CRISPR/Cas9 editing. That made us be interested in the investigation of exon skipping events systematically in large cancer cohorts' datasets from TCGA project and three other published studies. We identified this kind of events in 191 genes with the false positive rate estimated to be around 6%. Among these genes, we recaptured the well-known skipping event of exon 14 in the hepatocyte growth factor receptor (MET) in lung cancer patients. We also observed the same skipping events in both DNA mismatch repair gene MLH1 and renin receptor ATP6AP2 in lung and head and neck cancers although they were previously reported to be causative in inherited colorectal cancer and Parkinson disorder, respectively. In addition, we identified three kinds of novel skipping events of the same exon in the tumor suppressor PTEN in breast cancer. One of them might be able to produce an in frame protein with internal deletion of 128 amino acids affecting the phosphatase and catalytic domain. Most importantly, we discovered the gene fusion between cyclin dependent kinase inhibitor CDKN1A and RAS oncogene related protein RAB44. This gene fusion, accompanied by the exon skipping events within CDKN1A, was merely caused by a single nucleotide variation of a splice site of CDKN1A, contrary to the common knowledge that the gene fusions occur mainly as the result of large scale structural variations. Furthermore, the protein sequence of RAB44 was intact but the expression was activated clearly. Considering the high specificity that RAB44 is not expressed in all types of normal tissues and the relatively high prevalence of this kind of gene fusion in bladder cancer (1%), skin melanoma (1%) and stomach cancer (1/400), diagnosis of subgroups and targeted therapy must be worth further study.

Introduction

Exon skipping is the most common model of alternative splicing in mammalian cells[1]. However, it is rather rare to see a skipping of a constitutive exon, which consistently presents in all the known isoforms. We observed this kind of events when trying to knock out NGLY1 gene in a disease study about NGLY1 deficiency[2] through inducing a small deletion in its third exon using CRISPR/Cas9 technology. We expected this frame shift deletion would generate a premature termination codon (PTC) and cause the transcripts to be degraded by nonsense-mediated mRNA decay (NMD) pathway. But there was clear

evidence that the second and fourth exon was joined directly, which means the third exon was skipped in the mature mRNAs. How a small deletion that is far from the exon intron boundary causes the skipping event? The first case about the observation of exon skipping due to the mutation without changing the splice site was in *Dystrophin* in a Duchenne Muscular Dystrophy (DMD) patient in 1991[3]. After that, several other cases were anecdotally reported, including FBN1 in Marfan syndrome (MFS)[4], ADA in severe combined immunodeficiency disease (SCID)[5] and HMBS in acute intermittent porphyria (AIP)[6], etc. Valentine[7] summarized that those exon skipping events could either be explained by the disruption of exon splicing enhancer (ESE) or by the nuclear scanning mechanism accompanied by nonsense-mediated mRNA decay (NMD). Although the latter one was denied by both Liu et al. through generating nonsense and missense mutations in BRCA1 gene with the conclusion that nonsense mutations was neither necessary nor sufficient for the exon skipping[8] and Wang et al. who observed that the exon skipping wasn't affected when depleting UPF2, a key component of NMD pathway[9]. Actually, in addition to the conventional realization that the isoforms generated by exon skipping events, either through disruption of ESEs or the splice sites, might be able to have the effect of gain or loss of function, there are at least four additional benefits for comprehensive investigation of this kind of events. First, as illustrated in our NGLY1 example and in the study by Uddin et al.[10], we need to be more careful when designing gRNAs in the CRISPR/Cas9 experiments. Even there are not off targets effect of these gRNAs, the final RNAs or proteins may still not be our intendeds. Second, some hidden exon skipping causable mutations might be able to provide explanations for those severe mutations with ameliorated or no disease phenotype[11,12]. Third, the exon skipped proteins might be oncogenic drivers and attractive therapeutic targets, such as the skipping of exon 14 in hepatocyte growth factor receptor (MET) in non-small cell lung cancers[13,14] and the corresponding small molecular tyrosine kinase inhibitors including Crizotinib, Capmatinib and so on[15,16]. Forth, as a well-known example in the treatment of DMD, many more targetable sites might be explored when designing antisense oligonucleotide (AON)-mediated exon skipping therapy[17,18].

Basically, there are two ways to investigate the effect of mutations on exon skipping systematically. The first one is QTL mapping in large population. The individuals with the causative mutations may have significantly higher percentage spliced in (PSI) score[19-21] as mentioned by Monlong et al. in their splicing QTL package sQTLseekeR[22]. However, the majority of genetic mutations are rare and novel[23] that make QTL mapping, which usually requires at least 5 percent of the frequency, unable to detect most of the signals. The second is the one we proposed in this study through utilizing the somatic mutations in large cohorts of cancers. It is pretty straightforward that the only difference between tumor and tumor adjacent sample on the DNA level for a specific region is the somatic mutation itself. If we observed the skipping of an exon, which contains the somatic mutation, in the tumor sample but not in the tumor adjacent, it is most likely that this somatic mutation is responsible for the skipping as illustrated in Figure 1. In order to perform this type of analysis, there have to be both whole genome sequencing or whole exome sequencing (for the identification of somatic mutations) and RNA-Seq (for the identification of exon skipping events) for the matched tumor and tumor adjacent samples from the same patient. As shown in Supplementary Figure 1, in the TCGA project, there are 684 individuals among 23 different types of cancers meeting the above requirements. In addition, we also included 122 other patients of three cancers from published studies. With strict and conserved parameters, we finally identified exon skipping events in 191 unique genes. Permutation through random switching of the samples estimated that the false positive rate of our identification was between 3% and 11%, with the medium value as 6%. It is as expected that the majority of the skipping events were caused by mutations around the splice sites, but

other types of mutations (nonsense, missense, synonymous, etc.) might also have the effect most likely due the disruption of ESEs.

Among the 191 identified genes, we recaptured the skipping of exon 14 of MET. We also identified missense mutations and the corresponding skipping events in MLH1 and ATP6AP2, which were previously reported to be causative in inherited colon cancer[24] and Parkinson disorder[25,26], respectively. In addition, in a breast cancer patient, we observed three different kinds of novel skipping of exon 5 in the tumor suppressor PTEN. Two of them have frame shift effect while the third one might be able to produce an in frame protein with internal deletion of 128 amino acids. The most surprising finding in this study was about the gene fusion between CDKN1A and RAB44, accompanying with the skipping of the second exon of CDKN1A, due to the disruption of the splice site. This kind of fusion didn't change the protein sequence of RAB44 but activated it, which, as a RAS oncogene gene related protein, was not expressed in normal cells except in blood cells. It is well known that gene fusion is an important mechanism to activate oncogene and most of the fusion events are due to large structural variations (SVs), including translocation[27], deletion[28], etc., occurring on the DNA level. There are also several other cases talking about gene fusions on the RNA level caused by trans-splicing[29,30] or read through[31-33] events. However, to our knowledge, it is the first case reporting the gene fusion that was caused by SNVs or small INDELs around the splice site of the upstream fusion partner. Considering the high specificity that RAB44 is only expressed in the patients with this kind of fusion events and its relatively high prevalence in cancer cohorts (4 in 408 bladder cancer, 1 in 103 skin melanoma and 1 in 373 stomach cancer), diagnosis of subgroups and targeted therapy are worth further study.

Results

An anecdotal example of exon skipping in NGLY1 gene

In an attempt to knock out NGLY1 gene in a NGLY1 deficiency study, we designed a guide RNA targeting the internal region (62bp from the exon intron boundary) of the third exon (ENSE00003589640) in K562 cell line with the expectation that this frame shift small deletion created by the CRISPR/Ca9 system would generate a premature termination codon (PTC) and cause the transcripts to be degraded by the nonsense-mediated mRNA decay (NMD) pathway. Whole genome sequencing of the wild type and mutated cell lines confirmed the precise anchor of the target, without any obvious off target effects. However, a skipping event of the exon 3 was unexpectedly observed (Part A of Figure 2) in the RNA-Seq of the single clone cells (both clone 15 and clone 20). In order to quantify a skipping event, we introduced two variables which were defined by Wang et al.[21] and illustrated in Figure 1, including inclusive reads (IR) which indicates the number of reads supporting the inclusion of an exon and exclusive reads (ER) which measures the supporting reads of exclusion. As shown in Part B of Figure 2, it is pretty significant (p value < 2e-11) that the skipping event occurred in the mutated cells rather than wild type cells. We verified this event using RT-PCR (Part C of Figure 2) and Sanger sequencing. The third exon is a constitutive exon which presents in all the annotated isoforms in both ENSEMBL[34] and UCSC[35] database and it has never been observed to be skipped before (Part A of Figure 2). The skipping happens to be a kind of in frame translation, producing a 572aa protein which is 82aa shorter than the wild type one. It merits further study whether this shorter version of protein has any loss or gain of function or even dominant negative effects. However, it seems the PUB domain, a putative protein-protein interaction domain, has been partially deleted. Our NGLY1 example together with the findings from Uddin et al.[10] and many other anecdotally reported exon skipping cases remind us that we need to be very careful with the design of the guide RNAs in the CRISPR/Cas9 system.

Extensive identification of exon skipping events caused by somatic mutations in cancers

There were 684 individuals across 23 different types of cancers in the TCGA project sequenced on both DNA (whole exome) and RNA (RNA-Seq) level in the tumor tissue together with the matched tumor adjacent. In addition, 122 individuals from three other published studies, including 68 colon cancer[36], 32 gastric cancer[37] and 22 small cell lung cancer[38], with similar scenario were also included. First, somatic mutations were identified using the whole exome sequencing data with the comparison between tumor and the matched tumor adjacent samples. As shown in Supplementary Figure 1, though the number varies widely in different individuals and cancer types, one individual could generally carry hundreds of somatic mutations within the exon regions[39,40]. Second, as the basis of this study was the argument that a skipping event of a somatic mutation containing exon is most likely caused by the somatic mutation itself if this skipping only occurred in the tumor sample rather than the matched tumor adjacent, we performed the identification of exon skipping events using RNA-Seq data from the same individual (both tumor and tumor adjacent) focusing on those exons with somatic mutations discovered in the previous step. The quantification of a skipping event here mimics the in vitro case of NGLY1. The tumor samples are like the mutated cells generated by CRISPR/Cas9 system and the tumor adjacent samples could be regarded as wild type cells. Basically, we required the number of exclusive reads (ER) was large enough in the tumor but zero or almost zero in the tumor adjacent samples. In practice, as described in more details in the Methods section bellow, with strict filtering parameters, we finally identified exon skipping events in 191 unique genes (Supplementary Table 1). In order to estimate the false positive rate of our identification, we randomly switched the samples hundreds of times and found from 5 to 21 skipping events in each round, supporting the false positive rate should be between 3% and 11% with the medium value as 6% (Supplementary Figure 2). These false positive hits are not necessary meaning that they are all artificial, instead, it suggests we could observe the skipping event of an exon but couldn't find any somatic mutations there. Potential explanations include the missing identification of somatic mutations due to the low coverage of some specific exome regions or the complicated mixture of tumor and tumor adjacent cells, the unknown somatic mutations in the nearby intron regions and even the trans-regulation effects by other genes such as the deregulation from a RNA binding protein[19]. It is as expected that various kinds of mutations, including nonsense, missense and synonymous and so on, might be responsible for the exon skipping events, however, only the mutations disrupting splice sites are significantly enriched (Supplementary Figure 3).

In two lung cancer patients, we recaptured a skipping event of exon 14 in the hepatocyte growth factor receptor gene MET, which was caused by either a single nucleotide variation or a 6bp deletion nearby the splice site (Figure 3). This kind of skipping was first discovered by Ma et al. in a small cell lung cancer patient in 2003[41] and it has been supposed to disrupt ubiquitin-mediated degradation, leading to a relative increase of MET protein level, and contribute to the oncogenic activation[42]. It is an attractive target for the cancer therapy giving the prevalence of around 3% in lung cancers and at least five small molecular tyrosine kinase inhibitors (TKIs), including crizotinib, are being investigated clinically with promising preliminary results[15].

We identified a missense mutation (chr3: 37000961) and the corresponding skipping event of exon 3 in a DNA mismatch repair gene MLH1 in a small cell lung cancer (SCLC) patient. The exact same mutation and skipping event have already been reported in a patient with Lynch syndrome or hereditary non-polyposis colorectal cancer (HNPCC)[24]. There was evidence that this kind of defect of MLH1 was responsible for inherited colorectal cancer[41], although whether it is also able to contribute to SCLC merits

further study.

There was a missense mutation (chrX: 40597303) in a renin receptor gene ATP6AP2 on chromosome X that was 41bp away from the splice site but the skipping event of the mutation containing exon 4 was still detected. Studies of patient derived cells from the parkinsonism disorder has already showed that another missense mutation (chrX: 40597293, rs397518480) markedly increased the skipping of exon 4, resulting in significant overexpression of the exon skipping isoform that produces in frame protein with internal deletion of 32 amino acids and concomitant reduction of the normal isoforms containing this exon[25,26]. ATP6AP2 is an essential component of the vacuolar ATPase required for lysosomal degradation and autophagy, and the reduction of normal isoforms containing exon 4 may compromise the vacuolar ATPase function and ultimately be responsible for the pathology. It was a little bit surprising that both these two missense mutations, although 10bp away, could cause the skipping of the mutation containing exon because the usually defined exon skipping enhancer motif was only around 7bp.

We also identified a 38bp deletion near a splice site of a tumor suppressor gene PTEN, which is responsible for the skipping of exon 5 observed in a breast cancer patient. This skipping event was verified independently in a human leukemia T-cell line (PF-382) carrying a 4bp insertion around the same splice site. As illustrated in Figure 3, it is interesting that there are at least three kind of events observed about the skipping of exon 5, including the direct joint between exon 4 and exon 6 (ES-46), exon 3 and exon 6 (ES-36), and exon 4 and exon 7 (ES-47). It seems the skipping would more likely to choose to include the nearest exons although the nearby adjacent exons might also be utilized (Supplementary Figure 4). It is not clear whether one single cell could make up all these different kind of skipping events or different cell clones generate them separately. All these three events have neither been annotated by ENSEBL or UCSC database. ES-46 and ES-36 might have frame shift effect and make much short proteins when predicted based on the canonical sequence defined by Uniprot[43] database (P60484-1), while ES-47 has in-frame effect, which might produce a 276aa protein, leading to the deletion of amino acids from 85th to 212nd. It seems the first two isoforms are more likely to be degraded by pathways such as NMD, however, if the third isoform could really be translated and be able to exist stable, it probably has special effect due to the disruption of phosphatase and catalytic domain.

Gene fusion between CDKN1A and RAB44 activated the expression of RAB44

In addition to the exon skipping events within single gene mentioned above, we observed unexpectedly the fusion between cyclin dependent kinase inhibitor CDKN1A and RAS oncogene related protein RAB44, which was accompanied by the skipping of the second exon of CDKN1A (Figure 4, Supplementary Figure 4). As the start codons of both these two genes located in their second exon, this kind of gene fusion, which joined the coding region of RAB44 directly to the downstream of the UTR region of CDKN1A, didn't change the protein sequence of RAB44 at all. However, RAB44 was activated. As shown in Figure 5, RAB44 expressed highly in all the 6 cancer samples (4 bladder cancers, 1 stomach cancer, and 1 skin melanoma) with this kind of gene fusion, but not in either the tumor adjacent or 30 other randomly chosen cancer samples without the fusion events. Actually, RAB44 didn't express or expressed extremely low in all the human tissues except for blood cells (Supplementary Figure 5). After checking about the genotype of all these 6 patients with this gene fusion, we found 5 of them had the splice site of the second exon of CDKN1A mutated either by SNVs or small INDELs (Supplementary Figure 6). In addition, we couldn't detect any large scale structure variations or other recurrent somatic mutations in either the whole exome sequencing (WES) or RNA-Seq data of these samples. Thus we proposed a model to explain the activation of RAB44 in Figure 6. The disruption of the splice site of

CDKN1A induced the exon skipping events; in the meanwhile, the fusion between CDKN1A and the downstream RAB44 also occurred and the fusion activated the expression of RAB44.

However, there exists an alternative model because in our data, the exon skipping of CDKN1A and the gene fusion between CDKN1A and RAB44 always present simultaneously. The activation of RAB44 might also be caused by the down regulation of CDKN1A (due to the skipping of exon 2, Supplementary Figure 7), rather than the gene fusion between these two genes. We illustrated the alternative mode in Supplementary Figure 8. With the following two aspects carefully checked, we believe the alternative model is less likely. First, if the activation of RAB44 was really caused by the down regulation of CDKN1A, we should somehow be able to observe negative correlation between CDKN1A and RAB44 in large cohorts. As shown in Part A of Figure 7, we couldn't observe this kind of trend. Second, when we zoomed in our alignments of the RNA-Seq reads to the exon level, we couldn't see the first exon of RAB44 in all the six samples, supporting the expression of RAB44 came from the gene fusion events (Part B of Figure 7). Although we couldn't totally exclude the possibility that the first exon, especially when it is short, might not be easy to be detected itself in the RNA-Seq data. It is worth mentioning that, besides of the absent of the first exon, we noticed in our TCGA-BT-A42C (T2) sample, the second and third exon also couldn't be detected (Part B of Figure 7). Actually, this sample had extremely higher expression compared to the other 5 samples (Figure 5). As shown in Supplementary Figure 9, this sample carries different kind of fusion as before. It joins the forth exon of RAB44 directly to the first exon of CDKN1A and it is surprising that this kind of fusion, skipping of the first three exons, could still be able to have intact protein sequence, because RAB44 has another isoform really beginning from the forth exon (Figure 4). The 1bp insertion, which is 11bp away from the splice site, might be responsible for this kind of different fusion event, but we are not sure whether this mutation site still belongs to the scope of the splice site or there is a hidden exon splicing enhancer there.

We have already known all the samples with the splice site mutated have the fusion and skipping events. Furthermore, in all these samples RAB44 has been activated. However, we are also curious about how many of the RAB44 activated samples could be explained by the mutation or fusion and skipping events. As RAB44 doesn't express or expresses extremely low in human tissues except for blood cells (the expression of RAB44 in blood cells are not due to gene fusion events), we checked all the 10945 samples with RNA-Seq data from TCGA, excluding acute myeloid leukemia. As shown in Supplementary Figure 10, the majority of them (10930 samples, 99.86%) have less than 500 reads mapped and only 15 samples have over 500 reads each. All our 6 samples are among these 15 samples, including the highest sample as mentioned before (TCGA-BT-A42C). We could neither see any mutations around the splice site nor see any exon skipping or gene fusion events in the remaining 9 samples. It is not clear whether there is additional trans-regulation effect or it is just because we sequenced the mixer of blood cells with the tissue cells.

How the pathway perturbated and finally contributed to the carcinogenesis? Although in the lack of replicates, we identified a very potential target, MDMD2, when comparing the 6 samples with over 50 random chosen samples without any fusion or skipping events. Actually, there were 110 genes significantly deregulated after controlling the effect by different tissues using DESeq2[44]. MDM2 became our top one candidate not only because it was the second highest significant gene following RAB44, but also because it was the only one expressing extremely highest in the sample TCGA-BT-A42C, presenting the same pattern as RAB44 which also expressed highest in this sample (Supplementary Figure 11). MDM2 encodes an E3 ubiquitin-protein ligase, which can promote tumor formation by targeting TP53 for proteasomal degradation. However, whether the up regulation of MDM2 is due to the

activation of RAB44 or it is caused by the down regulation of CDKN1A is not clear, giving that there was evidence supporting at least MDM2 could be able to interact with CDKN1A[45]. Thus, it merits further study whether the carcinogenesis is mainly contributed by the activation of RAS related oncogene or by the inactivation of tumor suppressor TP53 or both.

Verification of exon skipping events in human cancer cell lines

Rather than inducing mutations artificially, cancer cell lines provide an easier and quicker way to check whether any of the exon skipping events identified in tissues could be verified independently. As shown in Supplementary Table 2, the mutations of 34 genes among the 191 genes above could be found in one or more human cancer cell lines (within the $\pm$2 bp window), but only 20 of them have been sequenced with RNA-Seq in cancer cell line encyclopedia project (CCLE)[46]. Although these cell lines might be originated from different tissues where we observed the skipping events, there were still 11 genes could be verified. In addition to PTEN and MET mentioned before, we would like to highlight the most frequently mutated example in TP53 bellowing.

We identified a single nucleotide mutation (chr17: 7675237) in a colon cancer patient and a 11bp deletion (chr17: 7675231-7675241) in a breast cancer patient. Both of these two mutations have changed the splice site (chr17: 7675236) and might be responsible for the exon skipping event we observed on the RNA level. There are 12 cell lines carrying mutations nearby this splice site ($\pm$2 bp) but the RNA-Seq data of only 6 of them are available. As shown in Supplementary Figure 12, we also randomly included 24 other cell lines without any mutations nearby the splice site as the control samples. It is significant (p value = 0.001) that we could detect exon skipping events in 5 of the 6 mutated cell lines while there wasn't any skipping detected in all the 24 control cell lines. In addition, there might be explanation for the glioblastoma cell line (SF126), with mutations but without skipping events observed, that TP53 was not expressed in brain. These cancer cell lines not only verified the exist of the skipping event independently but further more provided very strong evidence that the exon skipping event occurred most likely only when the splice site was disrupted.

It is worth mentioning that what we observed here in TP53 is not a typical exon skipping event. Instead of joining two nonadjacent exons from the same isoform, this event seems to be formed by switching from one isoform to another and continuing to utilize the 5' UTR of the latter as illustrated in Supplementary Figure 13. This skipping event is not reported in UCSC database, but it has already been annotated by ENSEMBL database as TP53-020 (ENST00000604348), which might be able to produce a much short protein (143aa). Whether this short protein has specific function merits further study, especially when considering that there are 2 patients and 12 cell lines carrying mutations near the splice site, while there isn't any mutations nearby reported in normal population in the ExAC database[23].

Verification of exon skipping events using CRISPR/Cas9 in human cell line

We are trying to use CRISPR/Cas9 to disrupt both the splice site and the site 11bp away in bladder cancer cell line (and normal bladder tissue cells), expecting to observe gene fusion and exon skipping events there. About the mechanism, we are also curious how this gene fusion occurred. As we imagining, the fusion or exon skipping should occur during the processing of mRNAs from pre-mature to mature. Gene should be independent unit at that moment. How two genes could be joined then? In the pre-mature state, the two genes have already been transcribed together? Hope the following experiments would answer our questions.

Methods

Samples and somatic mutations

Although more than 10 thousand (11, 607) samples were sequenced with RNA-Seq in TCGA project, there were only 1396 samples (698 pairs) sequenced in a tumor and tumor adjacent matched manner. And 684 of the 698 pairs of samples were also sequenced with whole exome capture to identify somatic mutations. According to the quality control requirements of TCGA project, the average coverage of bases within the targeted exome is 150X or greater and for RNA-Seq at least 150 million reads generated per sample. All the somatic mutations of these 684 pairs of samples were extracted from the mutation annotation format (MAF) files based on MuTect[47] pipeline. In addition, 122 pairs of samples from three other cancers (colon cancer, small cell lung cancer and gastric cancer) were also included and the somatic mutations were collected from the supplementary tables of each publication[36-38]. The mutation coordinates of these three studies were converted from GRCh37 to GRCh38 using liftOver[35].

Identification of exon skipping events from RNA-Seq data

Upon the approval from the data access committee from dbGaP[48] and Genetech, we downloaded the bam files of the 684 pairs of samples of TCGA project from Genomic Data Commons (GDC) database and the fastq files of the 122 pairs of samples from European Genome-phenome Archive (EGA)[49] database, respectively. The bam files from the TCGA samples have already been mapped to GRCh38 using STAR[50] by GDC, thus we also mapped the fastq files from EGA to GRCh38 using STAR ourselves with default parameters. The number of inclusive reads (IR) and exclusive reads (ER) were calculated with different anchor (10bp and 20bp) length for the splitting of the reads, for each somatic mutation containing exon in each sample based on only uniquely mapped reads. Fisher's exact test was employed to calculate the significance of skipping capability (measured by IR and ER) between tumor and tumor adjacent sample for each exon. Those skipping events with p value smaller than 0.01 were nominated but the following filtering criteria were also included. First, for the tumor adjacent sample, IR must be at least 20 and ER mustn't be larger than 1 (the gene was expressed in the tumor adjacent, but no exon skipping was observed). But for the matched tumor sample, ER must be at least 20 (exon skipping was observed in the tumor sample). Second, in order to reduce false positive hits, we required the skipping events must be supported by the reads with the breaking points as known exon intron boundaries. Third, we further filtered the skipping events that have already been annotated as known isoforms in the UCSC database[35]. We didn't filter our events using ENSEMBL[34] database, because there were lots of predicted and poorly annotated isoforms which were also very interesting and should be given attention further, such as the short protein in TP53 as we motioned before.

Discussion

We presented a systematical investigation of exon skipping events utilizing the somatic mutations in large cohorts of cancers. Although we identified exon skipping events in 191 unique genes in total, we were only able to show case by case for a few of them, including MET, MLH1, ATPA62, PTEN, TP53, CDKN1A and RAB44. The remaining is not necessary less important, and we expected some of them might also be able to have essential effect in some conditions, such as in rare diseases.

It is worth to mention that, through our analysis, we only focused on those exons, which carry somatic mutations, with the considering of the limitation of computational resources. Actually, in order to identify

skipping events for all the exons, we need to scan at least 30,0000 exons each sample for all the 806 paired samples, instead of scanning only about 350 exons each sample here. Another consideration is even we could manage scanning all the exons and identified many more exon skipping events, we couldn't tell whether they are real and how they occur, because we are lacking the power to explain trans effect under current design of our project. A big limitation of the short reads sequencing is we can only see the local view of our skipping or fusion events, supporting by the splitting of single reads or improper distance of paired end reads. With the development and their application of the third generation long reads sequencing, such as Oxford Nanopore or PacBio technologies, in RNA or cDNA, the full length isoforms would get much better resolution and understanding. On the other hand, single cell sequencing would help a lot, because sometimes we observed different kinds of exon skipping or gene fusion events in the same sample, and, with the pooled sequencing, we couldn't be able to distinguish whether different cell clones made these events or one particular cell could produce different kinds of events. In addition, it is also very interesting to systematically check whether we can find exon inclusion events in cancers caused by somatic mutations due to disruption of splicing silencers.

Regarding the major finding of this study about the gene fusion between CDKN1A and RAB44, we are pretty interested in the drug discovery targeting RAB44, which is very specifically expressed in the patients with the fusion events. However, as a preliminary attempt to identify known compounds capable to deregulate RAB44 using drug-repositioning strategy based on LINCS and CMap datasets[51,52], we unfortunately found RAB44 itself was even not included in the array of those experiments. Actually, as the low coverage in the whole exome sequencing data suggests, this gene was also not targeted and captured by TCGA project. Besides, BioGRID[53] database didn't include any protein interaction information for this gene yet and PDB[54] database didn't have its protein structure information. Furthermore, when searching it in PubMed, we only got one entry talking about various members of Rab GTPase family[55]. Really lots of work needs to be done for characterizing this gene and deciphering its role in carcinogenesis. We are curious about the clinical phenotype of these 6 patients. However, we were not able to make a conclusion, giving that only three of them have clinical information collected by TCGA project.

In our analysis, we integrated the data from both DNA and RNA level, but it is also valuable to explore protein information. For example, we would like to know whether the isoforms formed by the exon skipping or gene fusion events could be able to produce proteins and whether those proteins could be existing stable and have loss or gain of function. As the first attempt, we checked the mass spectrometry data resources from Clinical Proteomic Tumor Analysis Consortium (CPTAC) database[56] and found three datasets curated from published studies, including 62 samples from ovarian cancers[57], 36 samples from breast cancers[58] and 93 samples from colorectal cancers[59]. However, only 4 samples among them overlapped with our 806 tumor and tumor adjacent matched samples. Unfortunately, we didn't see any exon skipping events in these 4 samples in our upstream RNA-Seq analysis. We also tried to cancel our restriction about the paired samples, and we found suspicious skipping events in 4 additional genes (ROCK1, IPO8, TMEM260 and BNIP1). We expected to see junction peptides or frame shift peptides[60] in the proteomics datasets from the sample patients; however, we didn't see any signal there. In addition, we tried to do similar analysis in the proteomics datasets from cancer cell lines[61]. Additional exon skipping events from four genes, including TP53, were checked in the matched cell lines but without any meaningful success. In fact, we expected at least the junction peptide of TP53 should be presented because from the annotation of Ensembl database, the exon skipping event was able to produce a short protein. Besides of the explanation that the skipped isoforms might not be able to be translated or they

were degraded, it might also be due to the low coverage of current proteomics technology, especially when considering the skipped protein itself might be lower than normal version, making it harder to be detected.

Finally, we tried to check systematically whether there were more gene fusion events besides of CDKN1A-RAB44 in our cancer cohorts. Indeed, we found two more events, including the fusion between SUMO2 and HN1 and the fusion between CLTC and VMP1. However, we couldn't see obvious functions of these genes in carcinogenesis and it merits further study whether they have other special effects.

Competing interests

The authors declare no competing financial interests.

Acknowledgements

We would like to thank Sandra Clauder, Fan Zhou and Ying Liu for helpful discussion and encourage. We would like to thank TCGA Research Network and European Genome-phenome Archive (EGA) for sharing with us the RNA-Seq data of the cancer cohorts. We also would like to thank all those patients donating biopsy for supporting the cancer research.

References

1 Sammeth, M., Foissac, S. & Guigo, R. A general definition and nomenclature for alternative splicing events. *PLoS computational biology* **4**, e1000147, doi:10.1371/journal.pcbi.1000147 (2008).

2 Enns, G. M. *et al.* Mutations in NGLY1 cause an inherited disorder of the endoplasmic reticulum-associated degradation pathway. *Genet Med* **16**, 751-758, doi:10.1038/gim.2014.22 (2014).

3 Matsuo, M. *et al.* Exon skipping during splicing of dystrophin mRNA precursor due to an intraexon deletion in the dystrophin gene of Duchenne muscular dystrophy kobe. *J Clin Invest* **87**, 2127-2131, doi:10.1172/JCI115244 (1991).

4 Dietz, H. C. *et al.* The skipping of constitutive exons in vivo induced by nonsense mutations. *Science* **259**, 680-683 (1993).

5 Santisteban, I. *et al.* Three new adenosine deaminase mutations that define a splicing enhancer and cause severe and partial phenotypes: implications for evolution of a CpG hotspot and expression of a transduced ADA cDNA. *Hum Mol Genet* **4**, 2081-2087 (1995).

6 Llewellyn, D. H. *et al.* Acute intermittent porphyria caused by defective splicing of porphobilinogen deaminase RNA: a synonymous codon mutation at -22 bp from the 5' splice site causes skipping of exon 3. *J Med Genet* **33**, 437-438 (1996).

7 Valentine, C. R. The association of nonsense codons with exon skipping. *Mutat Res* **411**, 87-117 (1998).

8 Liu, H. X., Cartegni, L., Zhang, M. Q. & Krainer, A. R. A mechanism for exon skipping caused by nonsense or missense mutations in BRCA1 and other genes. *Nat Genet* **27**, 55-58, doi:10.1038/83762 (2001).

9 Wang, J., Chang, Y. F., Hamilton, J. I. & Wilkinson, M. F. Nonsense-associated altered splicing: a frame-dependent response distinct from nonsense-mediated decay. *Mol Cell* **10**, 951-957 (2002).

10 Uddin, B., Chen, N. P., Panic, M. & Schiebel, E. Genome editing through large insertion leads to the skipping of targeted exon. *BMC Genomics* **16**, 1082, doi:10.1186/s12864-015-2284-8 (2015).

11 Kowalewski, C. *et al.* Amelioration of junctional epidermolysis bullosa due to exon skipping. *Br J Dermatol* **174**, 1375-1379, doi:10.1111/bjd.14374 (2016).

12 Chen, R. *et al.* Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. *Nat Biotechnol* **34**, 531-538, doi:10.1038/nbt.3514 (2016).

13 Pilotto, S. *et al.* MET exon 14 juxtamembrane splicing mutations: clinical and therapeutical perspectives for cancer therapy. *Ann Transl Med* **5**, 2, doi:10.21037/atm.2016.12.33 (2017).

14 Heist, R. S. *et al.* MET Exon 14 Skipping in Non-Small Cell Lung Cancer. *Oncologist* **21**, 481-486, doi:10.1634/theoncologist.2015-0510 (2016).

15 Reungwetwattana, T., Liang, Y., Zhu, V. & Ou, S. I. The race to target MET exon 14 skipping alterations in non-small cell lung cancer: The Why, the How, the Who, the Unknown, and the Inevitable. *Lung Cancer* **103**, 27-37, doi:10.1016/j.lungcan.2016.11.011 (2017).

16 Drilon, A., Cappuzzo, F., Ou, S. I. & Camidge, D. R. Targeting MET in Lung Cancer: Will Expectations Finally Be MET? *J Thorac Oncol* **12**, 15-26, doi:10.1016/j.jtho.2016.10.014 (2017).

17 Fairclough, R. J., Wood, M. J. & Davies, K. E. Therapy for Duchenne muscular dystrophy: renewed optimism from genetic approaches. *Nat Rev Genet* **14**, 373-378, doi:10.1038/nrg3460 (2013).

18 Syed, Y. Y. Eteplirsen: First Global Approval. *Drugs* **76**, 1699-1704, doi:10.1007/s40265-016-0657-1 (2016).

19 Guo, W. *et al.* RBM20, a gene for hereditary cardiomyopathy, regulates titin splicing. *Nat Med* **18**, 766-773, doi:10.1038/nm.2693 (2012).

20 Schafer, S. *et al.* Alternative Splicing Signatures in RNA-seq Data: Percent Spliced in (PSI). *Curr Protoc Hum Genet* **87**, 11 16 11-14, doi:10.1002/0471142905.hg1116s87 (2015).

21 Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470-476, doi:10.1038/nature07509 (2008).

22 Monlong, J., Calvo, M., Ferreira, P. G. & Guigo, R. Identification of genetic variants associated with alternative splicing using sQTLseekeR. *Nat Commun* **5**, 4698, doi:10.1038/ncomms5698 (2014).

23 Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291, doi:10.1038/nature19057 (2016).

24 McVety, S., Li, L., Gordon, P. H., Chong, G. & Foulkes, W. D. Disruption of an exon splicing enhancer in exon 3 of MLH1 is the cause of HNPCC in a Quebec family. *J Med Genet* **43**, 153-156, doi:10.1136/jmg.2005.031997 (2006).

25 Korvatska, O. *et al.* Altered splicing of ATP6AP2 causes X-linked parkinsonism with spasticity (XPDS). *Hum Mol Genet* **22**, 3259-3268, doi:10.1093/hmg/ddt180 (2013).

26 Poorkaj, P. *et al.* A novel X-linked four-repeat tauopathy with Parkinsonism and spasticity. *Mov Disord* **25**, 1409-1417, doi:10.1002/mds.23085 (2010).

27 National Academy of Sciences. *Science* **132**, 1488-1501, doi:10.1126/science.132.3438.1488 (1960).

28 Tomlins, S. A. *et al.* Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310**, 644-648, doi:10.1126/science.1117679 (2005).

29 Li, H., Wang, J., Mor, G. & Sklar, J. A neoplastic gene fusion mimics trans-splicing of RNAs in normal human cells. *Science* **321**, 1357-1361, doi:10.1126/science.1156725 (2008).

30 Rickman, D. S. *et al.* SLC45A3-ELK4 is a novel and frequent erythroblast transformation-specific fusion transcript in prostate cancer. *Cancer Res* **69**, 2734-2738, doi:10.1158/0008-5472.CAN-08-4926 (2009).

31 Valentijn, L. J., Koster, J. & Versteeg, R. Read-through transcript from NM23-H1 into the neighboring NM23-H2 gene encodes a novel protein, NM23-LV. *Genomics* **87**, 483-489, doi:10.1016/j.ygeno.2005.11.004 (2006).

32 Varley, K. E. *et al.* Recurrent read-through fusion transcripts in breast cancer. *Breast Cancer Res Treat* **146**, 287-297, doi:10.1007/s10549-014-3019-2 (2014).

33 Nacu, S. *et al.* Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples. *BMC Med Genomics* **4**, 11, doi:10.1186/1755-8794-4-11 (2011).

34 Aken, B. L. *et al.* The Ensembl gene annotation system. *Database (Oxford)* **2016**, doi:10.1093/database/baw093 (2016).

35 Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res* **12**, 996-1006, doi:10.1101/gr.229102. Article published online before print in May 2002 (2002).

36 Seshagiri, S. *et al.* Recurrent R-spondin fusions in colon cancer. *Nature* **488**, 660-664, doi:10.1038/nature11282 (2012).

37 Liu, J. *et al.* Integrated exome and transcriptome sequencing reveals ZAK isoform usage in gastric cancer. *Nat Commun* **5**, 3830, doi:10.1038/ncomms4830 (2014).

38 Rudin, C. M. *et al.* Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. *Nat Genet* **44**, 1111-1116, doi:10.1038/ng.2405 (2012).

39 Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-421, doi:10.1038/nature12477 (2013).

40 Martincorena, I. & Campbell, P. J. Somatic mutation in cancer and normal cells. *Science* **349**, 1483-1489, doi:10.1126/science.aab4082 (2015).

41 Ma, P. C. *et al.* c-MET mutational analysis in small cell lung cancer: novel juxtamembrane domain mutations regulating cytoskeletal functions. *Cancer Res* **63**, 6272-6281 (2003).

42 Kong-Beltran, M. *et al.* Somatic mutations lead to an oncogenic deletion of met in lung cancer. *Cancer Res* **66**, 283-289, doi:10.1158/0008-5472.CAN-05-2749 (2006).

43 The UniProt, C. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **45**, D158-D169, doi:10.1093/nar/gkw1099 (2017).

44 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).

45 Zhang, Z. *et al.* MDM2 is a negative regulator of p21WAF1/CIP1, independent of p53. *J Biol Chem* **279**, 16000-16006, doi:10.1074/jbc.M312264200 (2004).

46 Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603-607, doi:10.1038/nature11003 (2012).

47 Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31**, 213-219, doi:10.1038/nbt.2514 (2013).

48 Tryka, K. A. *et al.* NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res* **42**, D975-979, doi:10.1093/nar/gkt1211 (2014).

49 Lappalainen, I. *et al.* The European Genome-phenome Archive of human data consented for

biomedical research. *Nat Genet* **47**, 692-695, doi:10.1038/ng.3312 (2015).

50 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).

51 Lamb, J. *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929-1935, doi:10.1126/science.1132939 (2006).

52 Duan, Q. *et al.* LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. *Nucleic Acids Res* **42**, W449-460, doi:10.1093/nar/gku476 (2014).

53 Stark, C. *et al.* BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* **34**, D535-539, doi:10.1093/nar/gkj109 (2006).

54 Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235-242 (2000).

55 Srikanth, S., Woo, J. S. & Gwack, Y. A large Rab GTPase family in a small GTPase world. *Small GTPases* **8**, 43-48, doi:10.1080/21541248.2016.1192921 (2017).

56 Edwards, N. J. *et al.* The CPTAC Data Portal: A Resource for Cancer Proteomics Research. *J Proteome Res* **14**, 2707-2713, doi:10.1021/pr501254j (2015).

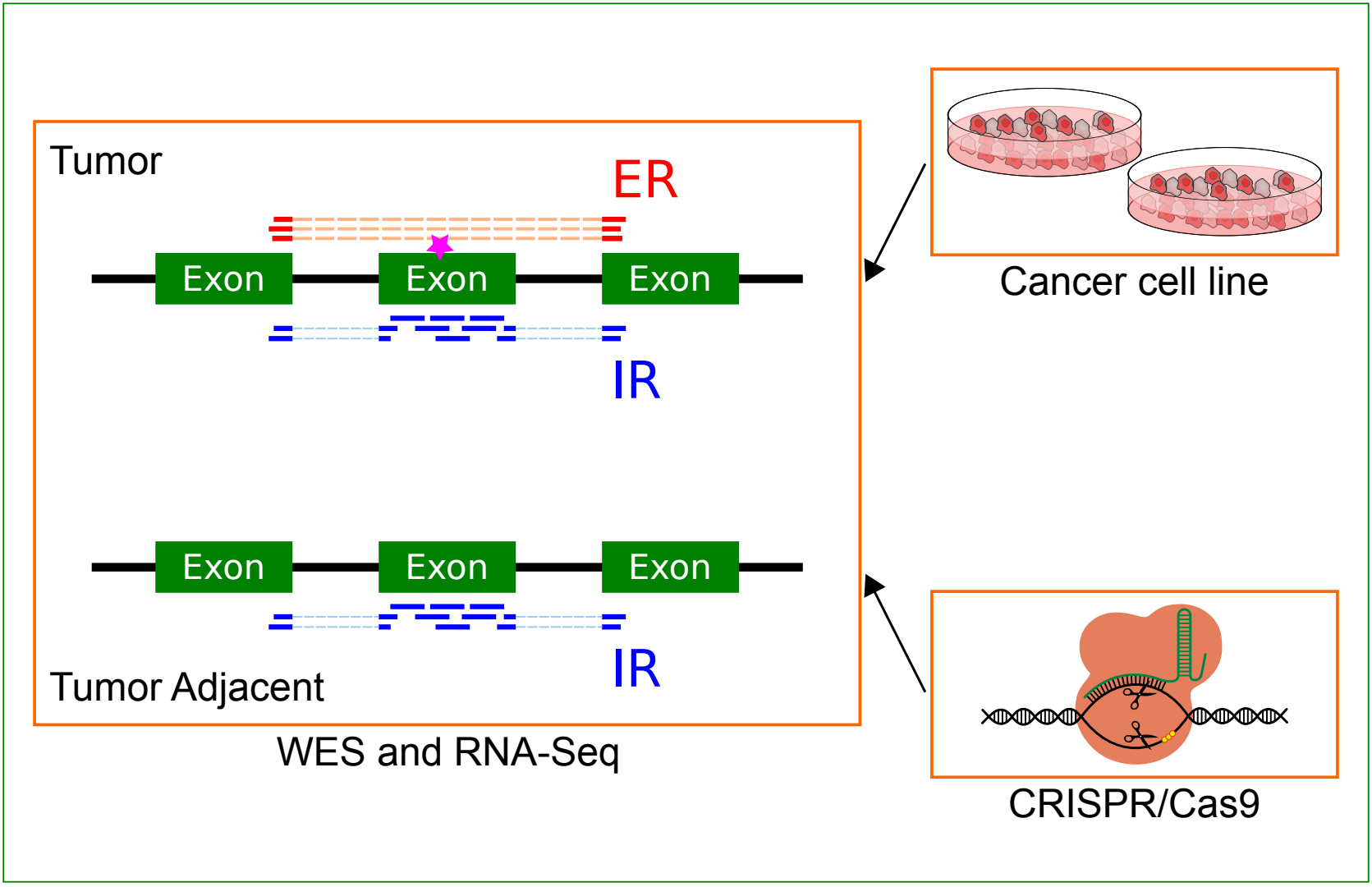
57 Zhang, H. *et al.* Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell* **166**, 755-765, doi:10.1016/j.cell.2016.05.069 (2016).

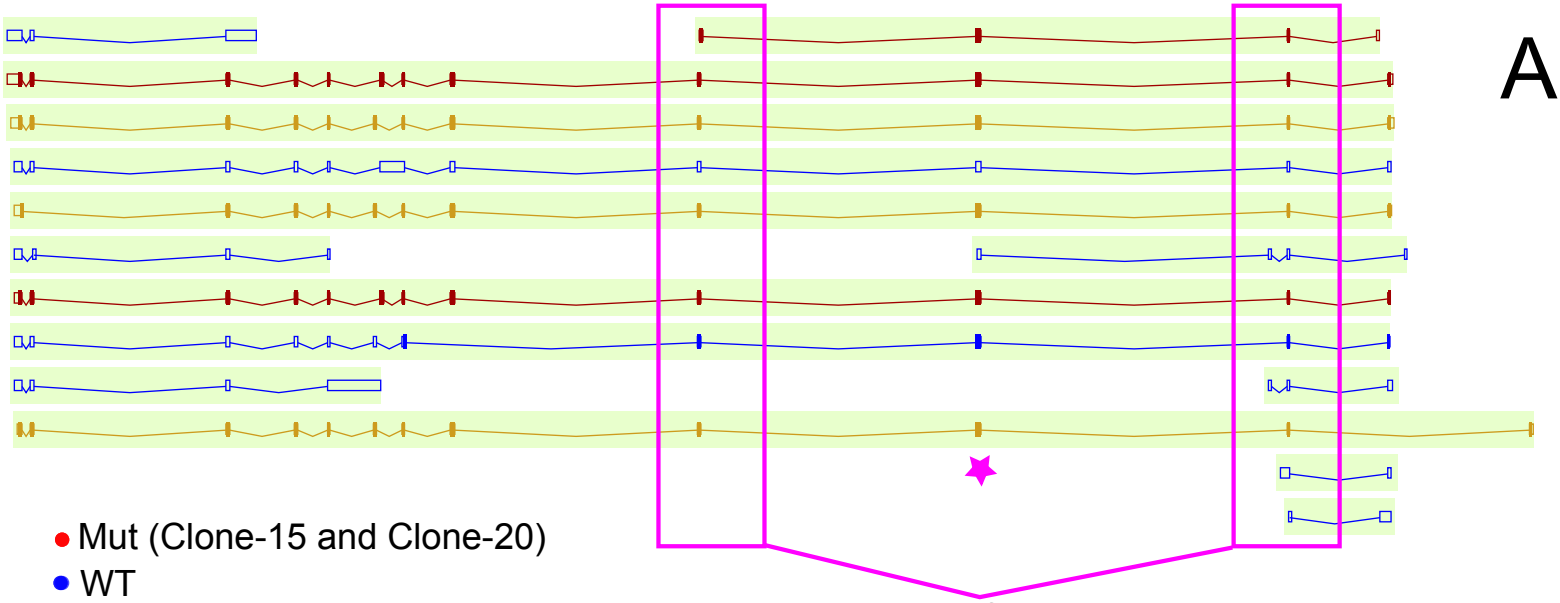
58 Mertins, P. *et al.* Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534**, 55-62, doi:10.1038/nature18003 (2016).

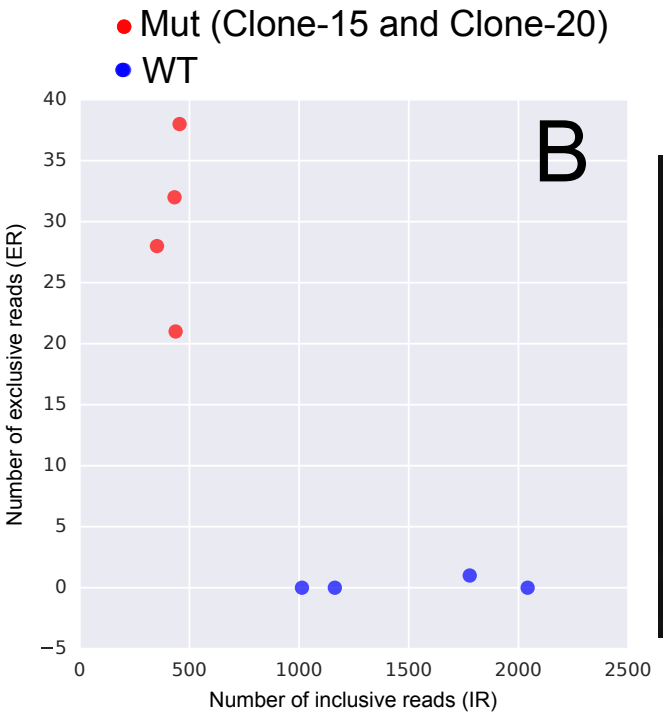
59 Zhang, B. *et al.* Proteogenomic characterization of human colon and rectal cancer. *Nature* **513**, 382-387, doi:10.1038/nature13438 (2014).

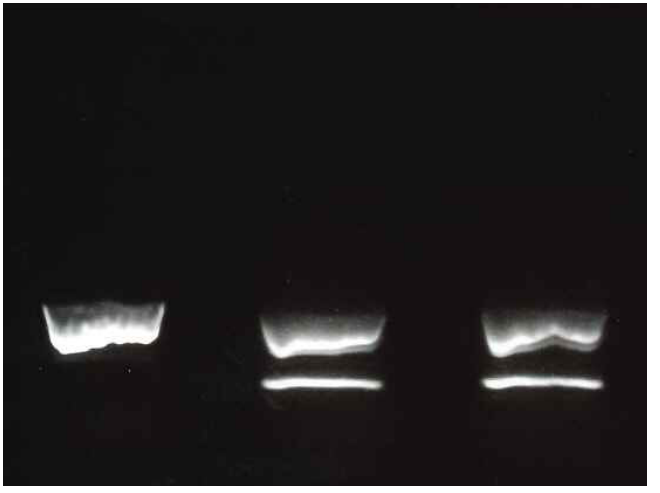
60 Sun, H. *et al.* Identification of HPV integration and gene mutation in HeLa cell line by integrated analysis of RNA-Seq and MS/MS data. *J Proteome Res* **14**, 1678-1686, doi:10.1021/pr500944c (2015).

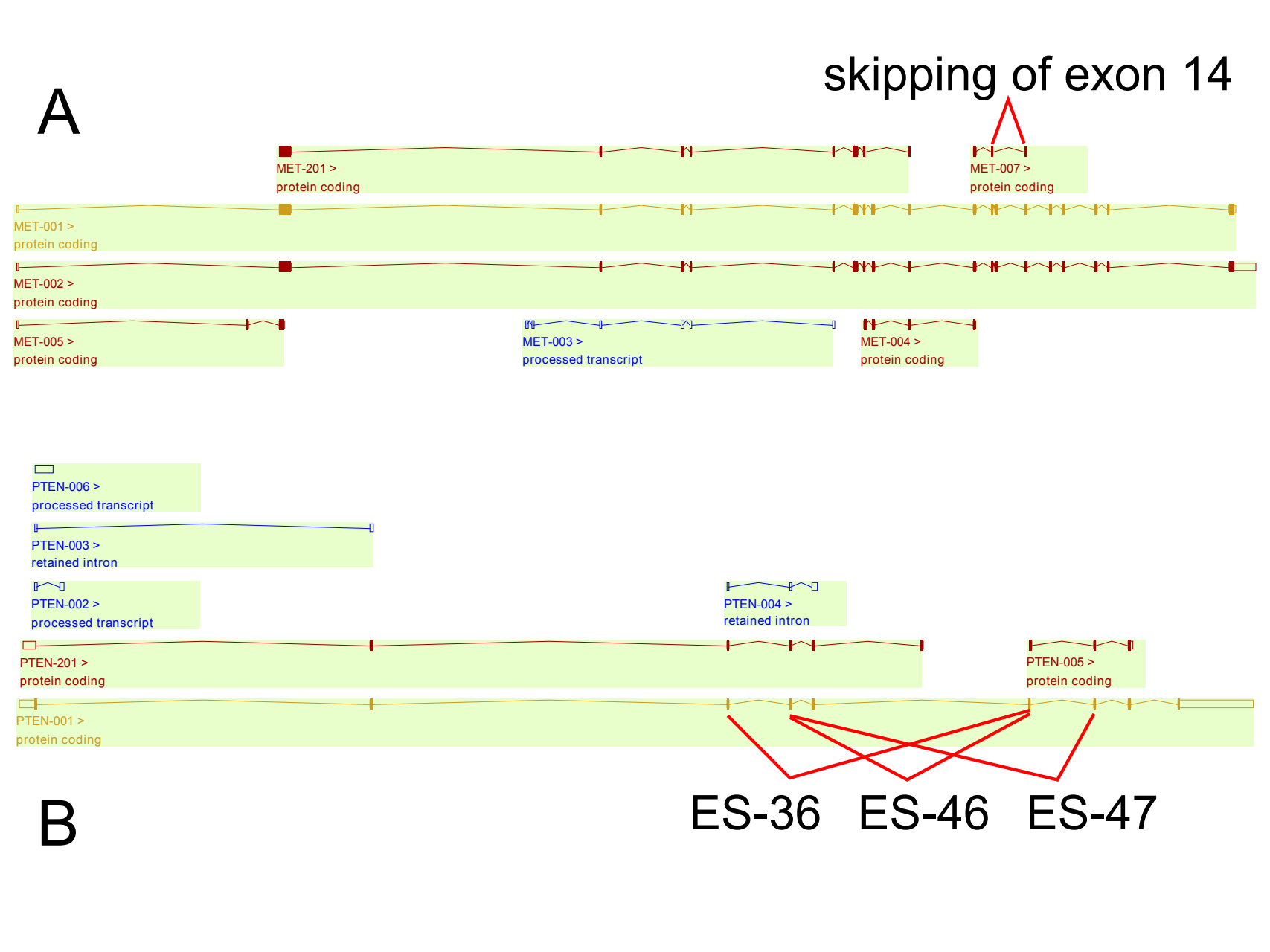
61 Gholami, A. M. *et al.* Global proteome analysis of the NCI-60 cell line panel. *Cell Rep* **4**, 609-620, doi:10.1016/j.celrep.2013.07.018 (2013).

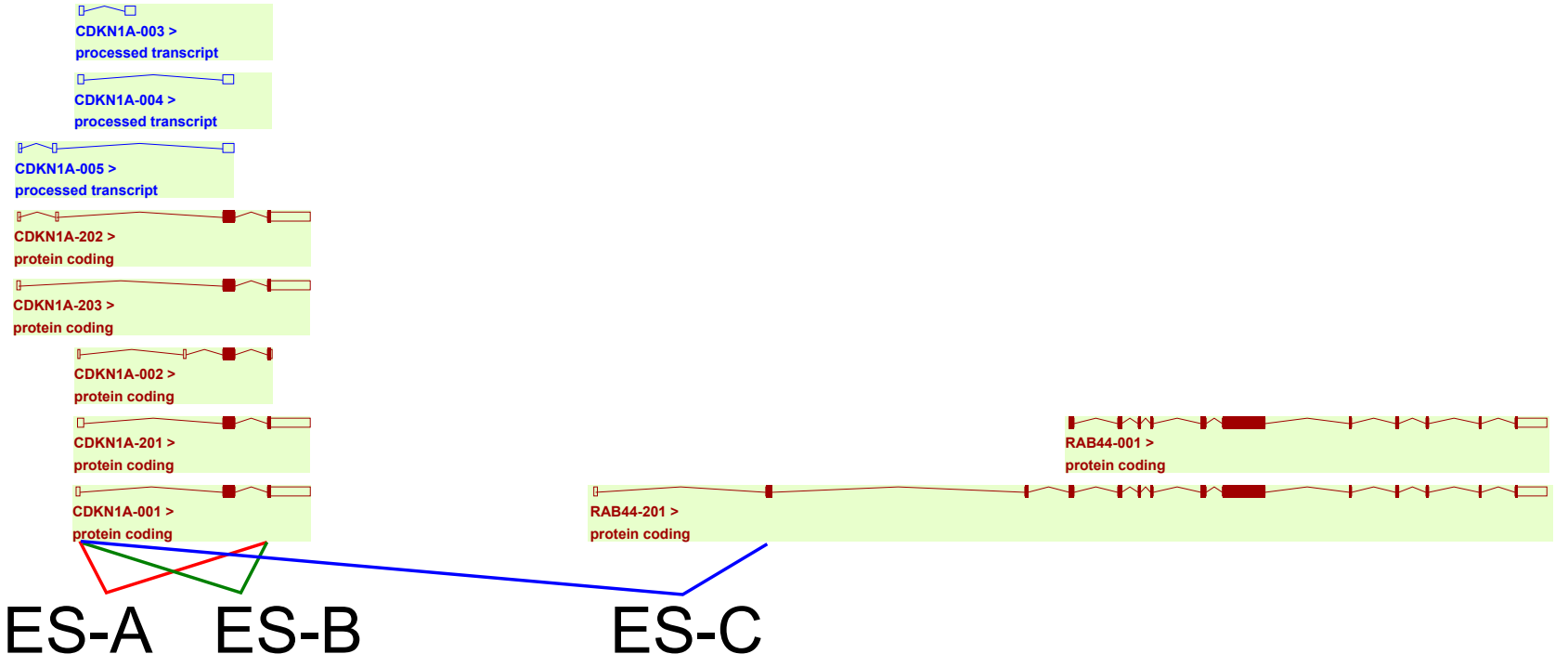

Tumor
ER
Exon
Exon
Exon
IR
Exon
Exon
Exon
IR
Tumor Adjacent
WES and RNA-Seq
Cancer cell line
CRISPR/Cas9

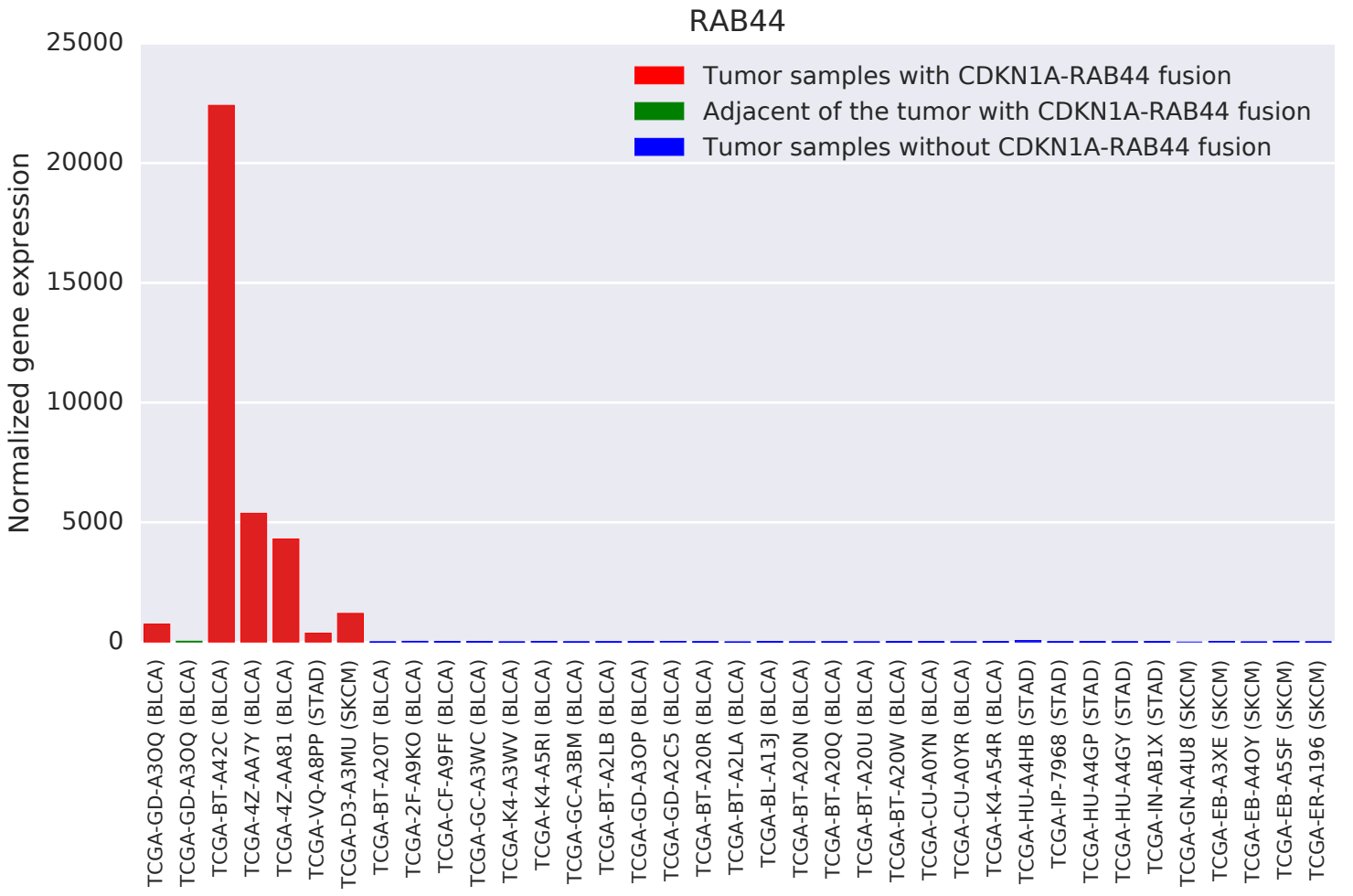

A
skipping of exon 3

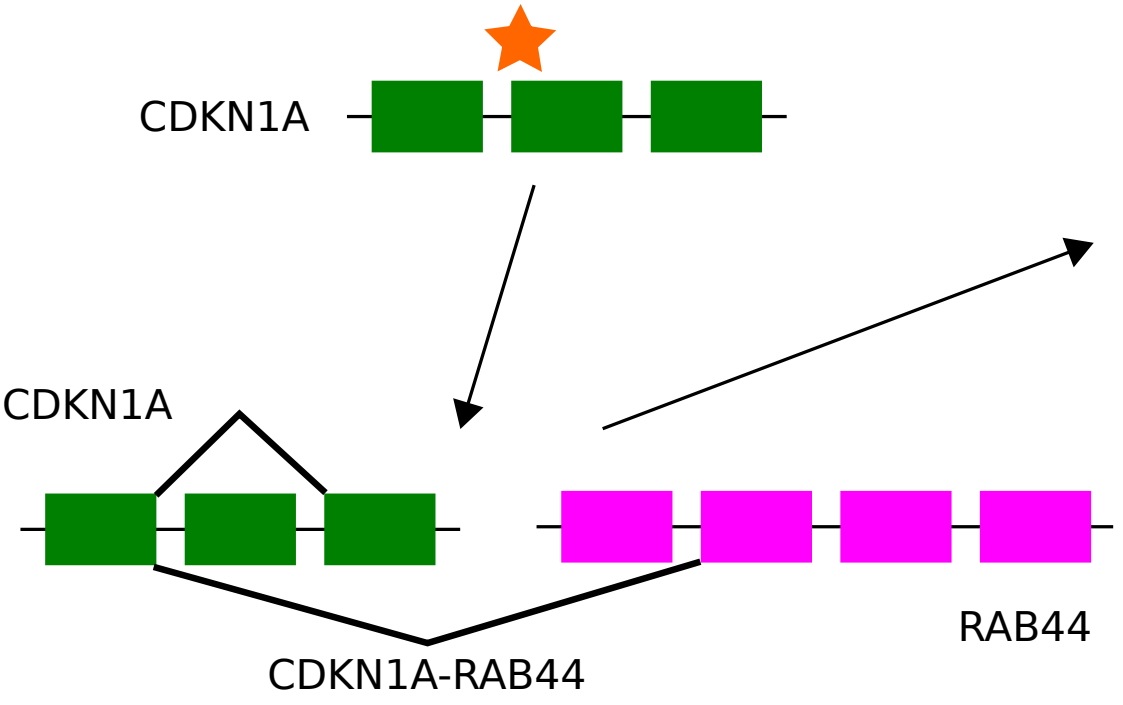

Mut (Clone-15 and Clone-20)
WT
B
Number of exclusive reads (ER)
Number of inclusive reads (IR)
40
35
30
25
20
15
10
5
0
−5
0
500
1000
1500
2000
2500

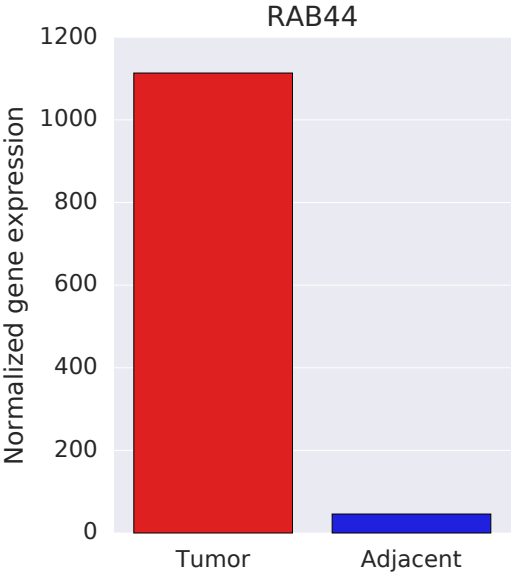

C
skipping of exon 3
WT
Clone-15
Clone-20

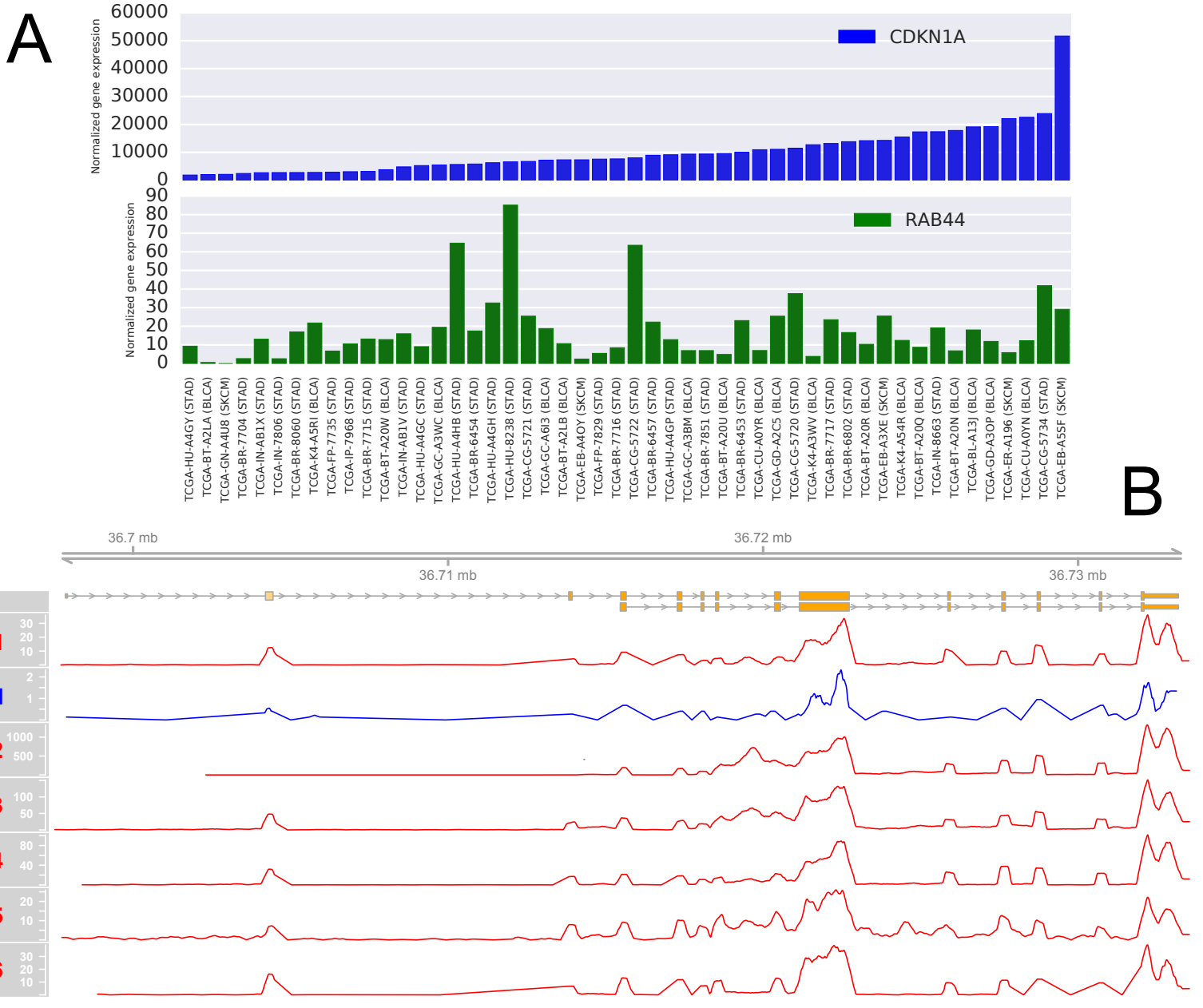

skipping of exon 14
A
MET-201 >
protein coding
MET-007 >
protein coding
MET-001 >
protein coding
MET-002 >
protein coding
MET-005 >
protein coding
MET-003 >
processed transcript
MET-004 >
protein coding
PTEN-006 >
processed transcript
PTEN-003 >
retained intron
PTEN-002 >
processed transcript
PTEN-004 >
retained intron
PTEN-201 >
protein coding
PTEN-005 >
protein coding
PTEN-001 >
protein coding
B
ES-36
ES-46
ES-47


CDKN1A-003 >
processed transcript
CDKN1A-004 >
processed transcript
CDKN1A-005 >
processed transcript
CDKN1A-202 >
protein coding
CDKN1A-203 >
protein coding
CDKN1A-002 >
protein coding
CDKN1A-201 >
protein coding
CDKN1A-001 >
protein coding
RAB44-001 >
protein coding
RAB44-201 >
protein coding
ES-A
ES-B
ES-C


RAB44
Tumor samples with CDKN1A-RAB44 fusion
Adjacent of the tumor with CDKN1A-RAB44 fusion
Tumor samples without CDKN1A-RAB44 fusion
Normalized gene expression
25000
20000
15000
10000
5000
0
TCGA-GD-A3OQ (BLCA)
TCGA-GD-A3OQ (BLCA)
TCGA-BT-A42C (BLCA)
TCGA-4Z-AA7Y (BLCA)
TCGA-4Z-AA81 (BLCA)
TCGA-VQ-A8PP (STAD)
TCGA-D3-A3MU (SKCM)
TCGA-BT-A20T (BLCA)
TCGA-2F-A9KO (BLCA)
TCGA-CF-A9FF (BLCA)
TCGA-GC-A3WC (BLCA)
TCGA-K4-A3WV (BLCA)
TCGA-K4-A5RI (BLCA)
TCGA-GC-A3BM (BLCA)
TCGA-BT-A2LB (BLCA)
TCGA-GD-A3OP (BLCA)
TCGA-GD-A2C5 (BLCA)
TCGA-BT-A20R (BLCA)
TCGA-BT-A2LA (BLCA)
TCGA-BL-A13J (BLCA)
TCGA-BT-A20N (BLCA)
TCGA-BT-A20Q (BLCA)
TCGA-BT-A20U (BLCA)
TCGA-BT-A20W (BLCA)
TCGA-CU-A0YN (BLCA)
TCGA-CU-A0YR (BLCA)
TCGA-K4-A54R (BLCA)
TCGA-HU-A4HB (STAD)
TCGA-IP-7968 (STAD)
TCGA-HU-A4GP (STAD)
TCGA-HU-A4GY (STAD)
TCGA-IN-AB1X (STAD)
TCGA-GN-A4U8 (SKCM)
TCGA-EB-A3XE (SKCM)
TCGA-EB-A4OY (SKCM)
TCGA-EB-A5SF (SKCM)
TCGA-ER-A196 (SKCM)


CDKN1A
CDKN1A
RAB44
CDKN1A-RAB44


RAB44
Normalized gene expression
1200
1000
800
600
400
200
0
Tumor
Adjacent


A
Normalized gene expression
CDKN1A
RAB44
TCGA-HU-A4GY (STAD)
TCGA-BT-A2LA (BLCA)
TCGA-GN-A4U8 (SKCM)
TCGA-BR-7704 (STAD)
TCGA-IN-AB1X (STAD)
TCGA-IN-7806 (STAD)
TCGA-BR-8060 (STAD)
TCGA-K4-A5RI (BLCA)
TCGA-FP-7735 (STAD)
TCGA-IP-7968 (STAD)
TCGA-BR-7715 (STAD)
TCGA-BT-A20W (BLCA)
TCGA-IN-AB1V (STAD)
TCGA-HU-A4GC (STAD)
TCGA-GC-A3WC (BLCA)
TCGA-HU-A4HB (STAD)
TCGA-BR-6454 (STAD)
TCGA-HU-A4GH (STAD)
TCGA-HU-8238 (STAD)
TCGA-CG-5721 (STAD)
TCGA-GC-A6I3 (BLCA)
TCGA-BT-A2LB (BLCA)
TCGA-EB-A4OY (SKCM)
TCGA-FP-7829 (STAD)
TCGA-BR-7716 (STAD)
TCGA-CG-5722 (STAD)
TCGA-BR-6457 (STAD)
TCGA-HU-A4GP (STAD)
TCGA-GC-A3BM (BLCA)
TCGA-BR-7851 (STAD)
TCGA-BT-A20U (BLCA)
TCGA-BR-6453 (STAD)
TCGA-CU-A0YR (BLCA)
TCGA-GD-A2C5 (BLCA)
TCGA-CG-5720 (STAD)
TCGA-K4-A3WV (BLCA)
TCGA-BR-7717 (STAD)
TCGA-BR-6802 (STAD)
TCGA-BT-A20R (BLCA)
TCGA-EB-A3XE (SKCM)
TCGA-K4-A54R (BLCA)
TCGA-BT-A20Q (BLCA)
TCGA-IN-8663 (STAD)
TCGA-BT-A20N (BLCA)
TCGA-BL-A13J (BLCA)
TCGA-GD-A3OP (BLCA)
TCGA-ER-A196 (SKCM)
TCGA-CU-A0YN (BLCA)
TCGA-CG-5734 (STAD)
TCGA-EB-A5SF (SKCM)
B
36.7 mb
36.71 mb
36.72 mb
36.73 mb
T1
A1
T2
T3
T4
T5
T6