Transcript 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[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

Alternative splicing, and specifically alternative exon inclusion, is one of the best studied mechanisms that create alternative isoforms. In its simplest form, alternative exon inclusion can be represented as exon skipping. Novel exon skipping events can be created through mutation events near and far from splice junctions, and thus may contribute to diseases, such as cancer. We systematically investigated exon skipping events in large cancer cohort datasets from The Cancer Genome Atlas (TCGA) project and three other published studies. We identified similar events in 191 genes with an estimated false positive rate of ~6%. Among these genes, we recaptured a known skipping event in exon 14 of the hepatocyte growth factor receptor (MET) in lung cancer patients. We also observed the similar exon skipping events in a DNA mismatch repair gene MLH1 and a renin receptor ATP6AP2 in lung and head and neck cancers. These exon skipping events were previously reported to be causative in inherited colorectal cancer and Parkinson disorder, respectively. In addition, we identified three novel exon skipping events of the same exon in the tumor suppressor PTEN in breast cancer. One of the events has the potential to produce an in frame protein with an internal deletion of 128 amino acids affecting the phosphatase and catalytic domain. Most importantly, we discovered a transcript fusion between cyclin dependent kinase inhibitor CDKN1A and RAS oncogene related protein RAB44. This transcript fusion, accompanied by the exon skipping events within CDKN1A, was caused only by a single nucleotide variation of a splice site of CDKN1A. Furthermore, the protein sequence of RAB44 was intact but its expression was clearly activated. The strong tissue specificity of RAB44 and the relatively high prevalence of this instance of transcript fusion in bladder cancer (1%), skin melanoma (1%) and stomach cancer (1/400) may warrant further study that could inform subclassifications of these cancers and the development of targeted therapies.

Introduction

Exon skipping is the most common mode of alternative splicing in mammalian cells[1]. One of the first cases of exon skipping was described due to mutations outside of a splice site in the *Dystrophin gene* in a Duchenne Muscular Dystrophy (DMD) patient in 1991[2]. After that, several other cases were reported, including FBN1 in Marfan syndrome (MFS)[3], ADA in severe combined immunodeficiency disease (SCID)[4], and HMBS in acute intermittent porphyria (AIP)[5]. Valentine[6] postulated that such exon skipping events could either be explained by the disruption of an exon splicing enhancer (ESE) or by the nuclear scanning mechanism accompanied by nonsense-mediated mRNA decay (NMD). However, the latter was contradicted by Liu et al., who generated nonsense and missense mutations in the BRCA1 gene to show that nonsense mutations are neither necessary nor sufficient for the exon skipping[7], as well as by Wang et al., who observed that the exon skipping was not affected when depleting UPF2, a key component of the NMD pathway[8]. Exon skipping events are important to investigate because they have the potential to cause gains or losses of function. Moreover, the skipping of exons containing severe disease mutation may be an explanation for individuals who bear these mutations yet display mild or no disease phenotypes[9,10]. The proteins produced from isoforms with skipped exons could 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[11,12] and the corresponding small molecular tyrosine kinase inhibitors including Crizotinib, Capmatinib[13,14]. Finally, as in the proposed treatments of DMD, many more targetable sites might be explored when designing antisense oligonucleotide (AON)-mediated exon skipping therapy[15,16].

Some efforts to systematically investigate the effects of mutations on exon skipping have utilized QTL mapping in large populations. The individuals with the causative mutations may have a significantly higher percentage spliced in (PSI) score[17-19] as described by Monlong et al. in their splicing QTL package sQTLseekeR[20]. However, the majority of genetic mutations are rare and novel[21] and with QTL mapping, which usually requires at least a 5 percent frequency, it is not possible to detect most of the signals. Another approach, which we use in the present study, investigates the effect of somatic mutations on exon skipping in large cancer cohorts (Figure 1). This analysis requires both whole genome/exome sequencing (for the identification of somatic mutations) and RNA-Seq (for the identification of exon skipping events) for matched tumor and tumor-adjacent samples from the same patient. The Cancer Genome Atlas (TCGA) datasets include 684 individuals with 23 types of cancers meeting our requirements. In addition, we also included 122 other patients of three cancers from published studies[22-24]. With strict and conserved parameters, we 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 median value as 6%. It is as expected that the majority of the skipping events were caused by mutations near or in splice sites, but other types of mutations (nonsense, missense, synonymous, etc.) were also associated with exon skipping events, likely due the disruption of ESEs.

Among the 191 identified genes involved in exon skipping was exon 14 of the MET gene. 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[25] and Parkinson disorder[26,27], 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 had frame shift effects, while the third would encode an in-frame protein with internal deletion of 128 amino acids. The most striking finding in our study was a transcript fusion between CDKN1A and RAB44, accompanied by the skipping of the second exon of CDKN1A due to a somatic mutation that disrupted its splice site. This transcript fusion did not change the protein sequence of RAB44, but activated its expression: as RAB44 is a RAS oncogene related protein, it would otherwise not be expressed in normal cells except in blood cells. It is well known that gene fusion is an important mechanism for activating oncogenes; most gene fusion events are due to large structural variations (SVs), including translocations[28], deletions[29], etc., occurring on the DNA level. There are also several reports of transcript fusions caused by trans-splicing[30,31] or read through[32-34] events. However, to our knowledge, the CDKN1A-RAB44 transcript fusion we report here is the first case caused by SNVs or small indels around the splice site of the upstream fusion partner. Considering that RAB44 is only expressed in patients with this transcript fusion event and the relatively high prevalence of this event in the cancer cohorts we examined (4 in 408 bladder cancer, 1 in 103 skin melanoma and 1 in 373 stomach cancer), this transcript fusion event may have implications for the clinical study of these cancer cases.

Results

Extensive identification of exon skipping events associated with somatic mutations in cancers

There were 684 individuals across 23 different types of cancers in the TCGA project with both DNA (whole exome) and RNA (RNA-Seq) sequences available from the tumor tissue along with the matched tumor adjacent. In addition, 122 individuals from three other published studies, including 68 colon cancer[24], 32 gastric cancer[22] and 22 small cell lung cancer[23] 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[35,36]. Second, as the basis of this study was the argument that a skipping event of a somatic mutation containing an 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). In order to quantify a skipping event, we introduced two variables which were defined by Wang et al.[19] 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 reads supporting exclusion. 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). 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, unknown somatic mutations in the nearby intron regions and even trans-regulation effects by other genes such as the deregulation from an RNA binding protein[17]. It is as expected that various kinds of mutations, (e.g., nonsense, missense, synonymous) 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 near the splice site (Figure 2). This kind of skipping was first discovered in a small cell lung cancer patient [37] and it is suspected to disrupt ubiquitin-mediated degradation, leading to a relative increase of MET protein level, and contribute to the oncogenic activation[38]. It is an attractive target for cancer therapy given the prevalence of this skipping event in ~3% of lung cancers and the fact that at least five small molecular tyrosine kinase inhibitors (TKIs), including crizotinib, are being investigated clinically to target the protein produced from this exon-skipping isoform with promising results[13].

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)[25]. There is evidence that this kind of defect in MLH1 is responsible for certain forms of inherited colorectal cancer[37], although whether it is also able to contribute to SCLC merits further study.

A missense mutation (chrX: 40597303) in a renin receptor gene ATP6AP2 on chromosome X was 41bp away from the splice site, but a skipping event associated with the mutation in exon 4 was detected. Studies of patient derived cells from the Parkinson's disease have already shown that another missense mutation (chrX: 40597293, rs397518480) can markedly increase the skipping of exon 4, resulting in significant overexpression of the exon skipping isoform producing an in frame protein with an internal deletion of 32 amino acids and concomitant reduction of the normal isoforms containing this exon[26,27]. 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.

We also identified a 38bp deletion near a splice site of the tumor suppressor gene PTEN, which is correlated and likely 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 2, it is interesting that at least three kinds of events were observed around the skipping of exon 5, including the direct connection of 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 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 perform all these different kind of skipping events or whether different clones would generate them separately. None of these three events have been annotated in either the ENSEMBL or UCSC databases. ES-46 and ES-36 might have a frame shift effect and make much shorter proteins than predicted based on the canonical sequence defined by the Uniprot[39] database (P60484-1). ES-47 has in-frame effects, which might produce a 276aa protein, leading to the deletion of amino acids 85 to 212. It seems the first two isoforms are more likely to be degraded by pathways such as NMD, however, if the third isoform is translated and stable, it could potentially exert a dominant negative effect due to the disruption of the phosphatase and catalytic domains.

Transcript fusion between CDKN1A and RAB44 activated the expression of RAB44

In addition to the exon skipping events within single genes mentioned above, we observed a novel transcript fusion between cyclin dependent kinase inhibitor CDKN1A and RAS oncogene related protein RAB44 transcripts, which was accompanied by the skipping of the second exon of CDKN1A (Figure 3, Supplementary Figure 4). As the start codons of both these two genes located in their second exon, this kind of transcript fusion, which joined the coding region of RAB44 directly to the downstream of the UTR region of CDKN1A, did not affect the protein sequence of RAB44. However, expression of RAB44 was activated by this fusion event. As shown in Figure 4, RAB44 was highly expressed in all 6 cancer samples with this transcript fusion (4 bladder cancers, 1 stomach cancer, and 1 skin melanoma), but not in the tumor-adjacent or 30 other randomly chosen cancer samples without the fusion events. RAB44 was either not expressed or expressed at a low level in all human tissues except for blood cells (Supplementary Figure 5). The genotype data of the 6 patients with this transcript fusion showed that in 5 of them, the splice site of the second exon of CDKN1A was mutated either by SNVs or small INDELs

(Supplementary Figure 6). In addition, no large scale structural variations or other recurrent somatic mutations were detected in either the whole exome sequencing (WES) or RNA-Seq data of these samples. To explain the activation of RAB44 in Figure 5, we propose that 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.

In the datasets we analyzed, the exon skipping of CDKN1A and the transcript fusion between CDKN1A and RAB44 always present simultaneously. This means that 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 transcript fusion between these two genes. We illustrated this alternative model in Supplementary Figure 8. The alternative model is less likely for two reasons: First, if the activation of RAB44 was really caused by the down regulation of CDKN1A, there should be an observable negative correlation between CDKN1A and RAB44 in large cohorts. As shown in Part A of Figure 6, this trend was not observed. Second, the alignments of the RNA-Seq reads at the exon level do not show the first exon of RAB44 in all the six samples, supporting the possibility that the expression of RAB44 came from the transcript fusion events (Part B of Figure 6). It cannot be excluded that the first exon, especially when short, might not be detected in the RNA-Seq data. It is worth noting that, in addition to the absence of the first exon in the TCGA-BT-A42C (T2) sample, the second and third exon also were not detected (Part B of Figure 6). Actually, this sample had much higher expression compared to the other 5 samples (Figure 4). As shown in Supplementary Figure 9, this sample displays a different fusion event where the fourth exon of RAB44 is spliced directly to the first exon of CDKN1A. This isoform results in an intact protein sequence because RAB44 has another isoform beginning from the fourth exon (Figure 3). The 1bp insertion, which is 11bp away from the splice site, might be responsible for this second type of fusion event, but we are not sure whether this mutation affects splicing or if there may be an exon splicing enhancer there.

All the samples with a mutated splice site display these fusion and skipping events. Furthermore, in all these samples RAB44 expression has been activated. We wondered how many of the RAB44 activated samples could be explained by the mutation or fusion and skipping events. As RAB44 is only expressed at low levels in human tissues except for blood cells, we analyzed all 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 6 transcript fusion samples are among these 15 samples. No mutations around the splice site nor any exon skipping or transcript fusion events were observed in the remaining 9 samples.

To investigate how these fusion events may contribute to carcinogenesis, we compared the 6 fusion samples to over 50 randomly chosen samples without any fusion or skipping events. Although these samples lack replicates, this comparison revealed 110 genes significantly deregulated after controlling for tissue specific effects using DESeq2[40]. Among these, MDM2 is an interesting candidate not only because it was the second highest significant gene following RAB44, but also because it was the only one expressed at a very high level in the sample TCGA-BT-A42C, just like RAB44 (Supplementary Figure 11). MDM2 encodes an E3 ubiquitin-protein ligase, which can promote tumor formation by targeting TP53 for proteasomal degradation. However, whether the upregulation of MDM2 is due to the activation of RAB44 or the downregulation of CDKN1A is not clear. There is evidence supporting the interaction between MDM2 and CDKN1A[41]. Further study will be required to determine whether carcinogenesis in these samples was influenced by the activation of the RAS related oncogene, the inactivation of tumor suppressor TP53, or both.

Verification of exon skipping events in human cancer cell lines

As shown in Supplementary Table 2, mutations in 34 genes among the 191 genes exhibiting exon skipping were found in one or more human cancer cell lines (within the ±2 bp window), but only 20 of them have been sequenced with RNA-Seq in the cancer cell line encyclopedia project (CCLE)[42]. Although these cell lines may have originated from different tissues where we observed the skipping events, there were still 11 genes that could be verified. In addition to PTEN and MET noted earlier, the most frequently mutated transcript was p53.

We identified a single nucleotide mutation (chr17: 7675237) in a colon cancer patient and an 11bp deletion (chr17: 7675231-7675241) in a breast cancer patient. Both of these mutations affect the splice site (chr17: 7675236) and may be responsible for the exon skipping event we observed on the RNA level. There are 12 cell lines carrying mutations near this splice site (±2 bp) but RNA-Seq data is only available for 6 of them. As shown in Supplementary Figure 12, we also randomly included 24 other cell lines without any mutations near the splice site as control samples. We detected significant exon skipping events (p value = 0.001) in 5 of the 6 mutated cell lines, versus no detected skipping in all 24 control cell lines. These cancer cell lines not only independently verified the existence of the skipping event but also provided evidence that the exon skipping event occurred most likely only when the splice site was disrupted.

It is worth noting 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 the UCSC database, but is in the ENSEMBL database as TP53-020 (ENST00000604348). Whether the truncated protein (143aa) encoded by this transcript has a functional impact merits further study, especially considering that there are 2 patients and 12 cell lines carrying mutations near the splice site, while no such mutations have been observed in control populations according to the ExAC database[21].

Methods

Samples and somatic mutations

Although more than 10,000 (11, 607) samples were profiled with RNA-Seq in the TCGA project, there were only 1396 samples (698 pairs) sequenced in a matched tumor and tumor-adjacent manner. 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, the average coverage of bases within the targeted exome is 150X or greater and for RNA-Seq at least 150 million reads were generated per sample. Somatic mutations of these 684 pairs of samples were extracted from the mutation annotation format (MAF) files based on the MuTect[43] 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[22-24]. The mutation coordinates of these three studies were converted from GRCh37 to GRCh38 using liftOver[44].

Identification of exon skipping events from RNA-Seq data

Upon approval from dbGaP[45] and Genetech, we downloaded the bam files of the 684 pairs of TCGA samples from the Genomic Data Commons (GDC) database and the fastq files of the 122 pairs of samples from European Genome-phenome Archive (EGA)[46] database, respectively. TCGA samples have already been mapped to GRCh38 using STAR[47] by GDC, thus we also mapped the fastq files from EGA to GRCh38 using STAR ourselves with default parameters. The numbers of inclusive reads (IR) and exclusive reads (ER) were calculated with different anchor (10bp and 20bp) lengths 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 capacity (measured by IR and ER) between tumor and tumor adjacent sample for each exon. Those skipping events with p values 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). For the matched tumor sample, ER had to be at least 20 (exon skipping was observed in the tumor sample). Second, in order to reduce false positive hits, we required that the skipping events 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[44]. We did not filter our events using the ENSEMBL[48] database.

Discussion

We have presented a systematic investigation of exon skipping events utilizing somatic mutations in large cancer cohorts. Our major finding concerning the transcript fusion of CDKN1A and RAB44, which may have implications in carcinogenesis and/or drug discovery. Little is known about RAB44: it has no known interactors and its structure has not yet been determined[49] [50]. [51]. Further research is needed to characterize this gene and decipher its role in carcinogenesis.

In our analysis, we integrated DNA and RNA data, but it would also be interesting to incorporate protein data. For example, it would be important to determine whether the isoforms formed by exon skipping or transcript fusion events produces protein and whether those proteins are stable or functional. Mass spectrometry data resources from Clinical Proteomic Tumor Analysis Consortium (CPTAC) database[52] contain three datasets curated from published studies, including 62 samples from ovarian cancers[53], 36 samples from breast cancers[54] and 93 samples from colorectal cancers[55]. However, only 4 samples among them overlapped with our 806 tumor and tumor adjacent matched samples, and we did not observe any exon skipping events in these 4 samples in our RNA-Seq analysis. We observed potential skipping events in 4 additional genes (ROCK1, IPO8, TMEM260 and BNIP1), but did not find junction peptides or frame shift peptides[56] in the patient proteomics datasets. Similar junction peptide analysis in the proteomics datasets from cancer cell lines[57] for four genes, including TP53 in the matched cell lines was also without success. This may be due to the skipped isoforms not being translated, being degraded, or due to insufficient coverage in the proteomics datasets.

Our study indicates that exon skipping and transcript fusion may be more prevalent consequences of somatic mutations in cancer than previously appreciated. Further work will be required to validate the impact of these events on protein expression and cellular function and what mechanistic role they may be playing in carcinogenesis or other disorders.

Competing interests

The authors declare no competing financial interests.

Acknowledgements

We would like to thank Sandra Clauder, Francesca Briganti, Fan Zhou and Ying Liu for helpful discussions. We would like to thank TCGA Research Network and European Genome-phenome Archive (EGA) for allowing us to analyze whole exome and transcriptome data from the cancer cohorts. The data/analyses presented in the current publication have been deposited in and are available from the dbGaP database under study accession phs000178.v9.p8 and EGA database under study accessions EGAS00001000288, EGAS00001000334 and EGAS00001000736.

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 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).

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

4 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).

5 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).

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

7 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).

8 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).

9 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).

10 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).

11 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).

12 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).

13 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).

14 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).

15 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).

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

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

18 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).

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

20 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).

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

22 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).

23 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).

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

25 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).

26 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).

27 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).

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

29 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).

30 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).

31 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).

32 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).

33 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).

34 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).

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

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

37 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).

38 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).

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

40 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).

41 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).

42 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).

43 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).

44 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).

45 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).

46 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).

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

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

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

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

51 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).

52 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).

53 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).

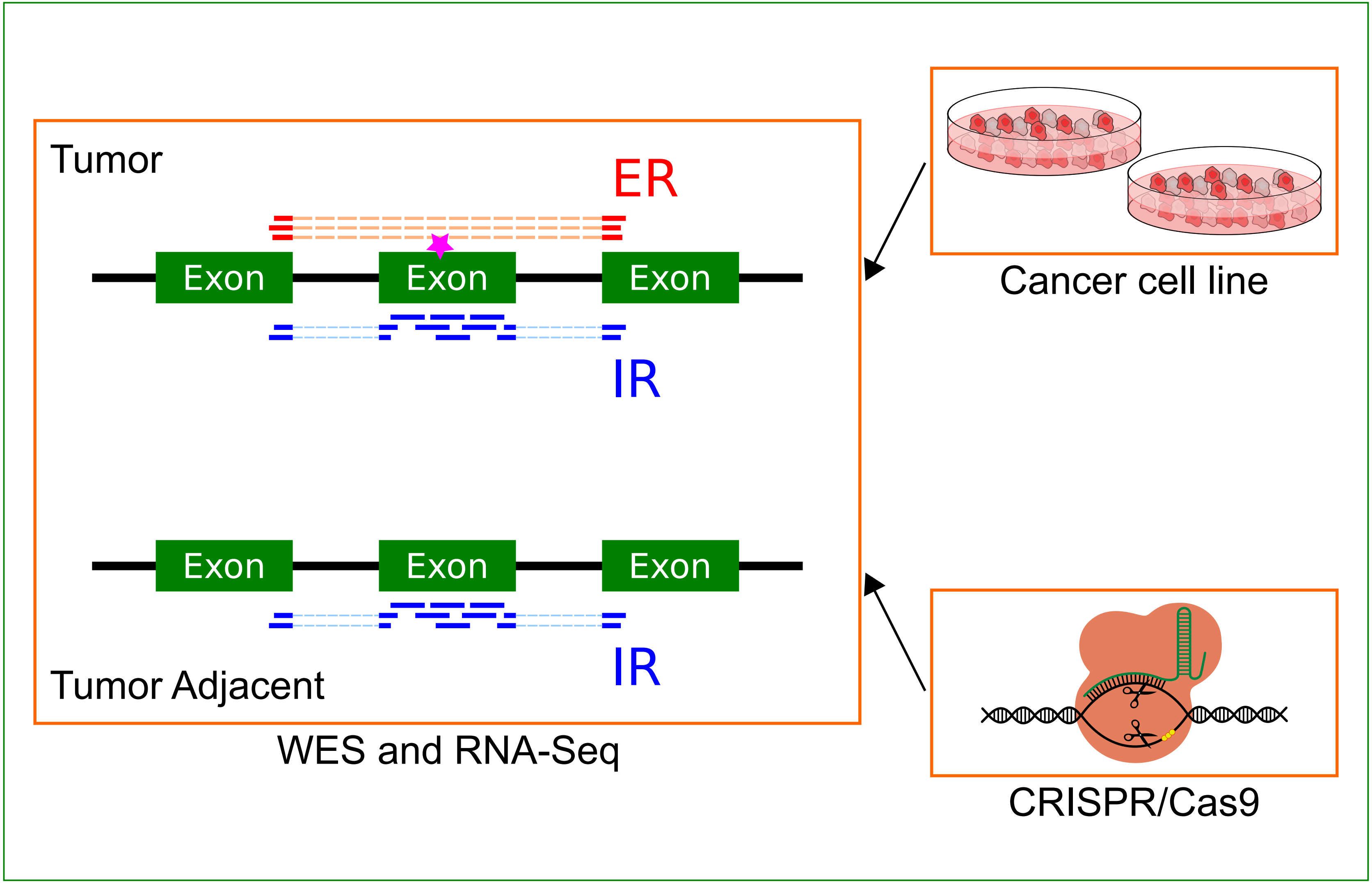
54 Mertins, P. *et al.* Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534**, 55-62, doi:10.1038/nature18003 (2016).

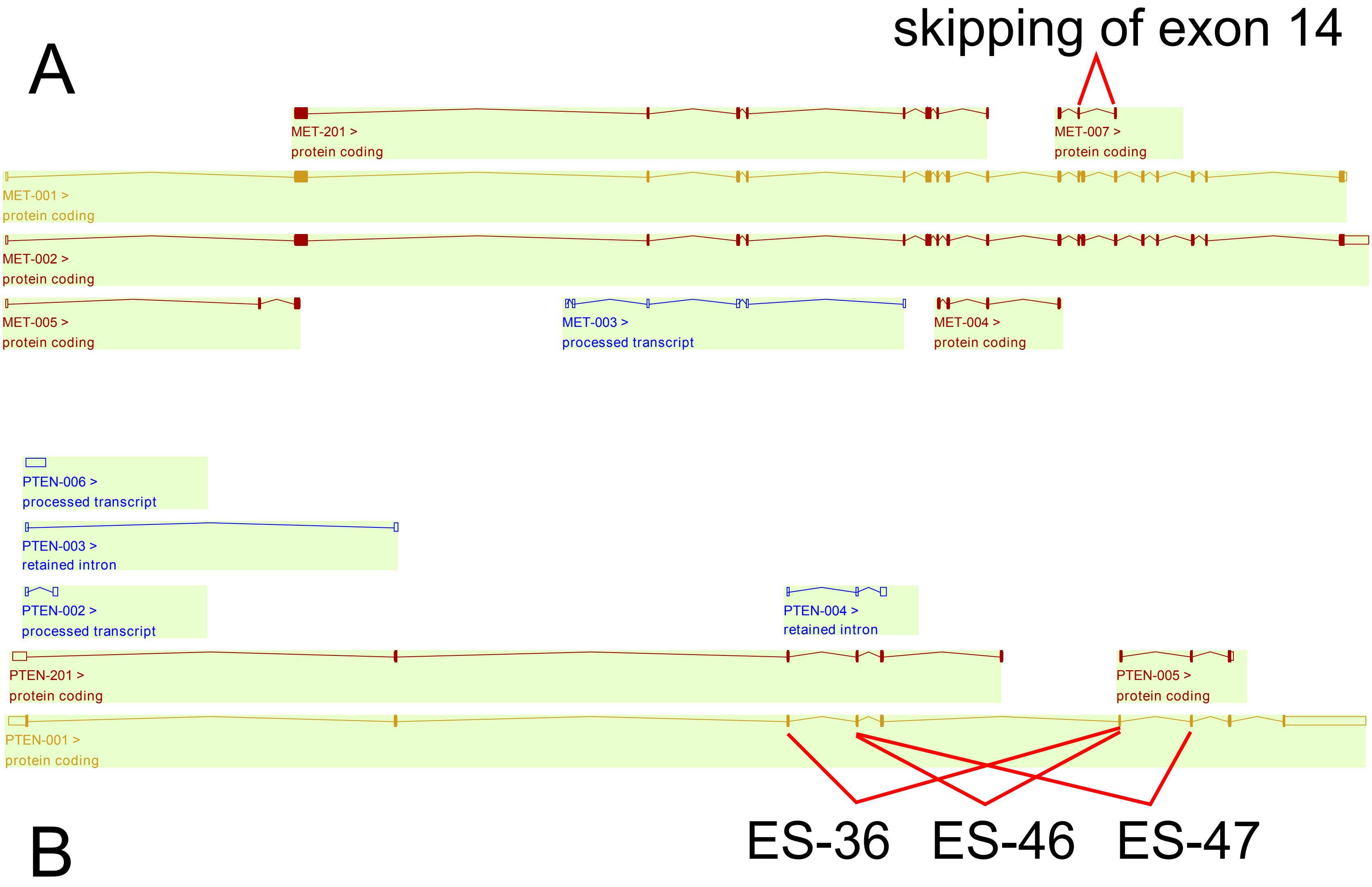
55 Zhang, B. *et al.* Proteogenomic characterization of human colon and rectal cancer. *Nature* **513**, 382-387, doi:10.1038/nature13438 (2014).

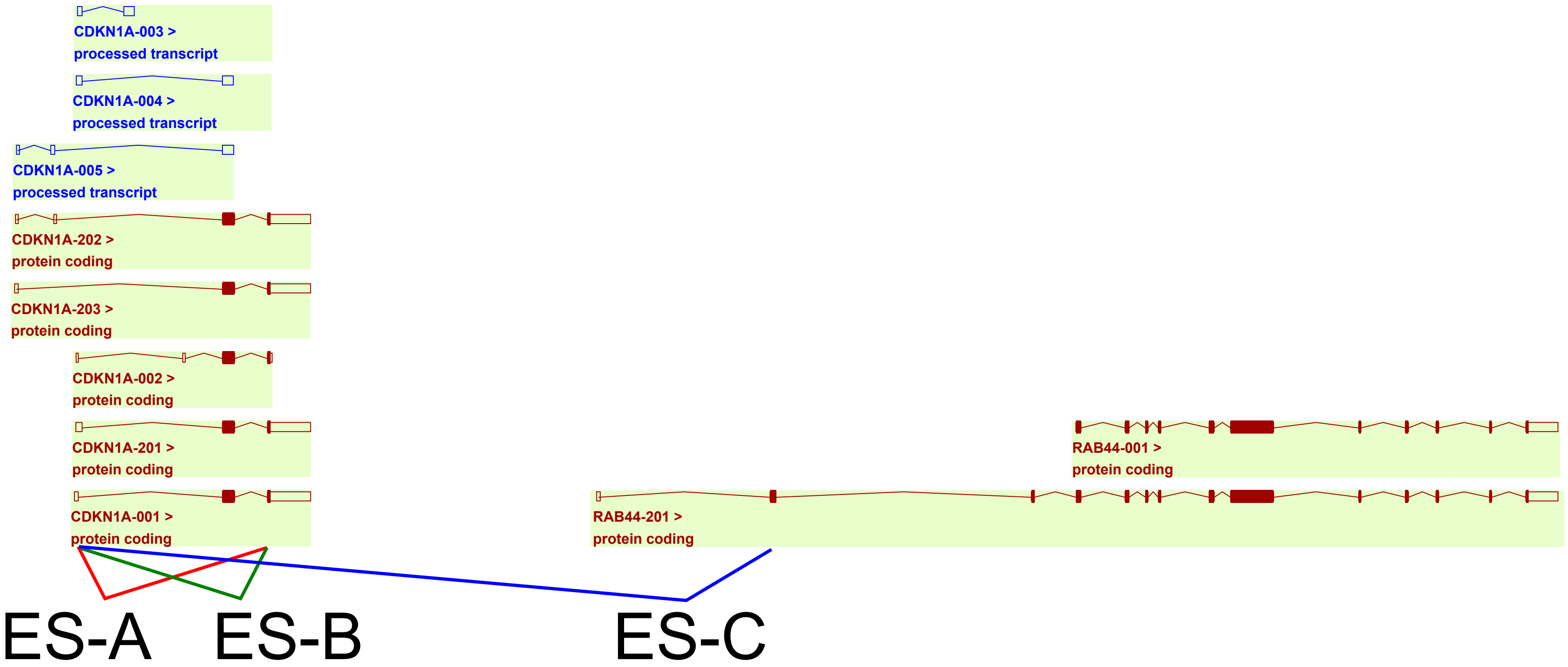
56 Sun, H. *et al.* Identification of HPV integration and gene mutation in HeLa cell line by

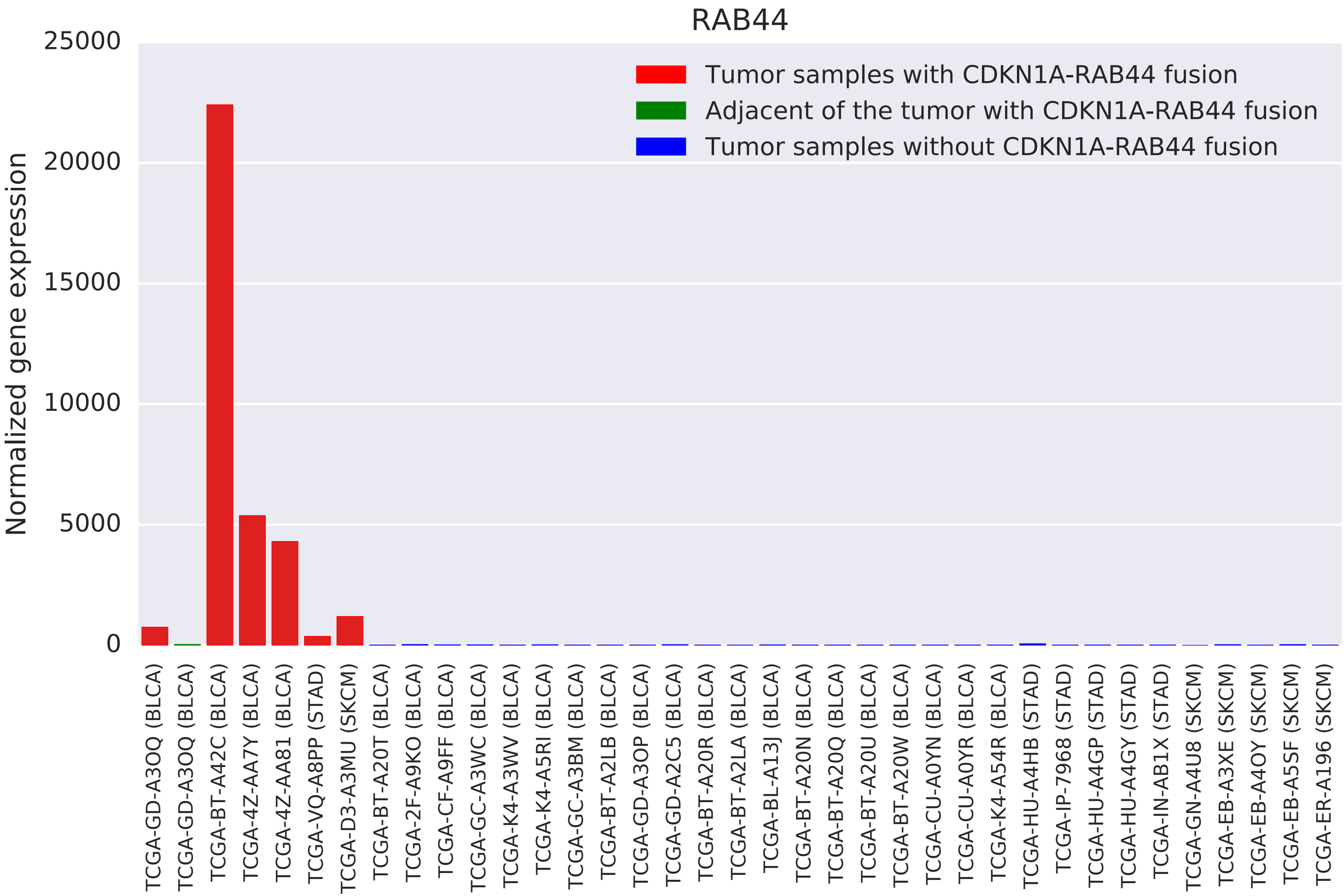
integrated analysis of RNA-Seq and MS/MS data. *J Proteome Res* **14**, 1678-1686, doi:10.1021/pr500944c (2015).

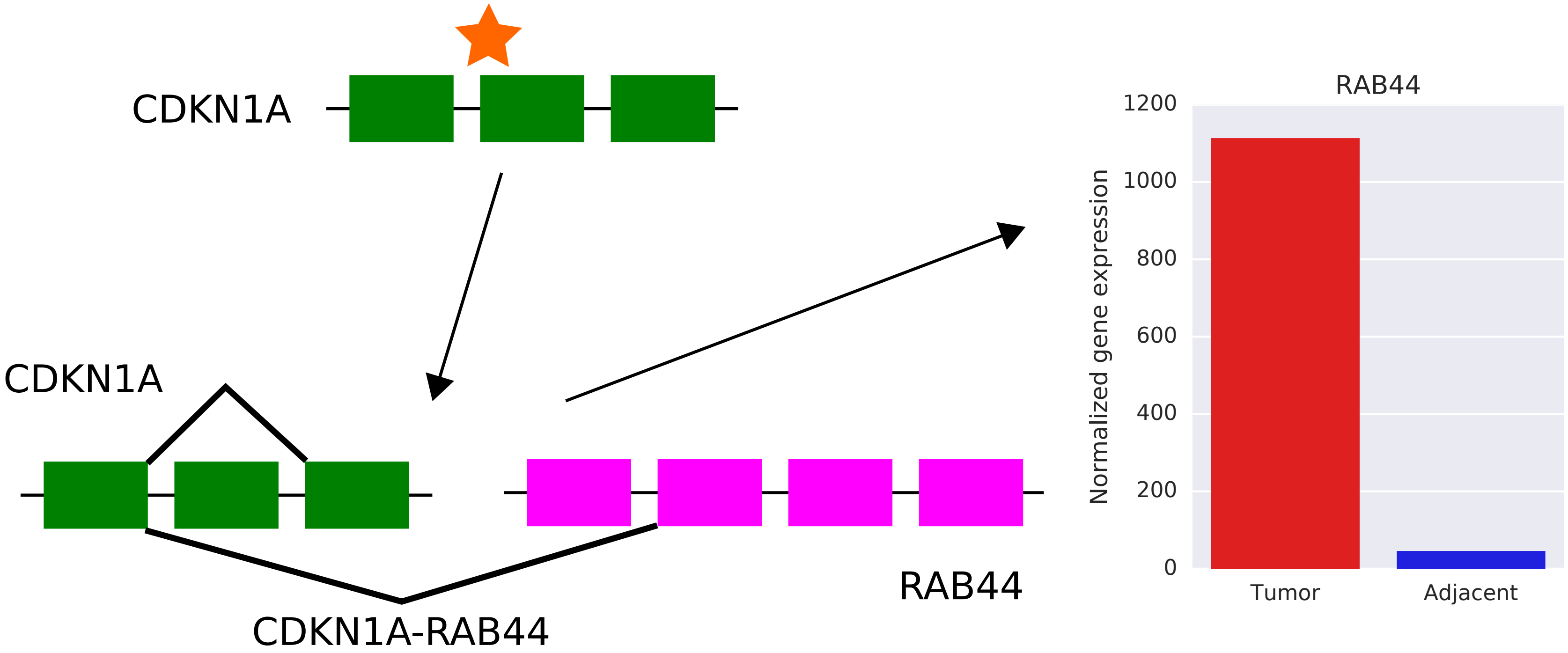
57 Gholami, A. M. *et al.* Global proteome analysis of the NCI-60 cell line panel. *Cell Rep* **4**, 609-620, doi:10.1016/j.celrep.2013.07.018 (2013).

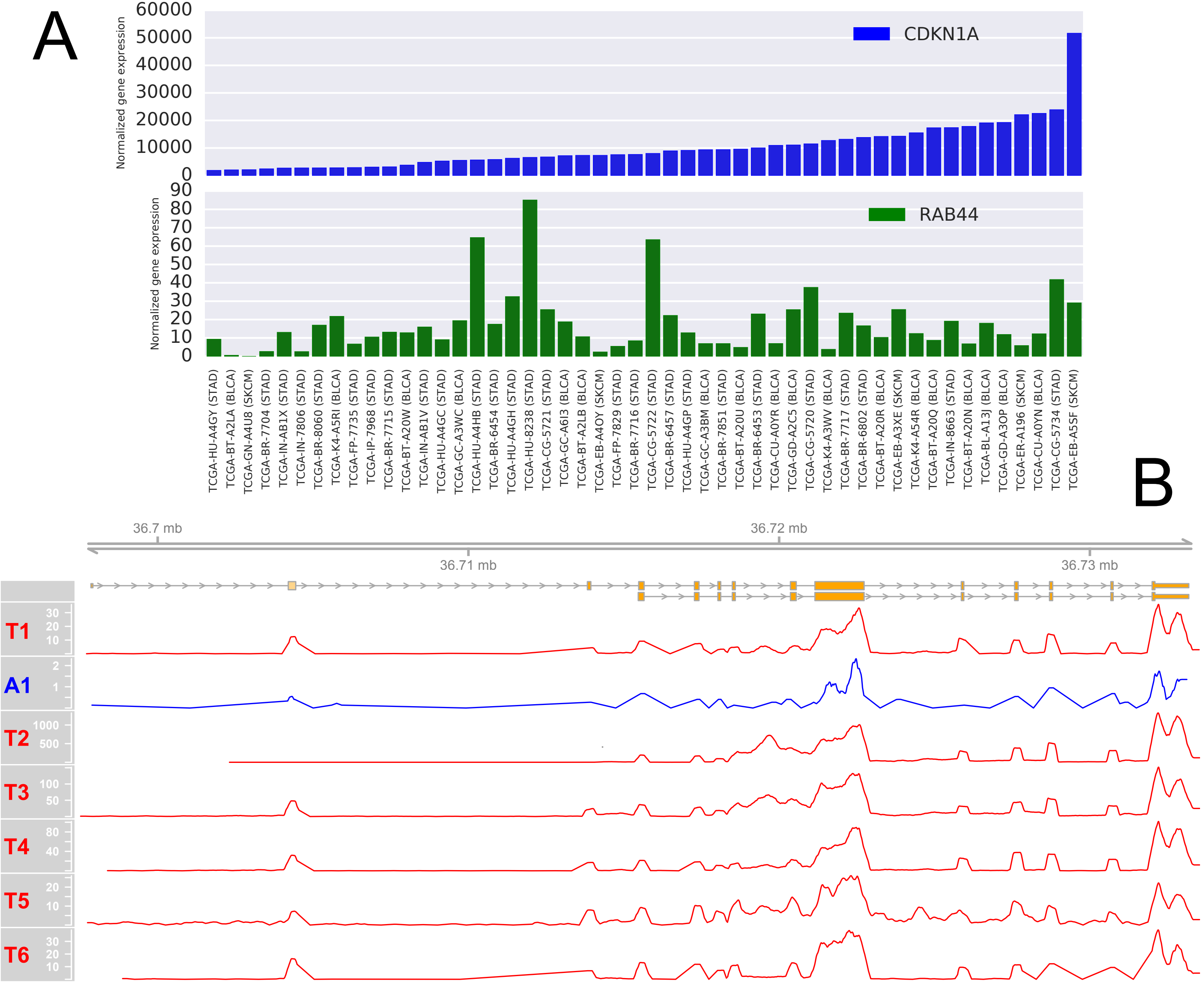

Tumor
ER
Exon
Exon
Exon
IR
Exon
Exon
Exon
IR
Tumor Adjacent
WES and RNA-Seq
Cancer cell line
CRISPR/Cas9


A
skipping of exon 14
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
RAB44-001 >
protein coding
CDKN1A-001 >
protein coding
RAB44-201 >
protein coding
ES-A
ES-B
ES-C


RAB44
Normalized gene expression
25000
20000
15000
10000
5000
0
🟥 Tumor samples with CDKN1A-RAB44 fusion
🟩 Adjacent of the tumor with CDKN1A-RAB44 fusion
🟦 Tumor samples without CDKN1A-RAB44 fusion
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
CDKN1A-RAB44
RAB44
RAB44
Normalized gene expression
1200
1000
800
600
400
200
0
Tumor
Adjacent


A
Normalized gene expression
60000
50000
40000
30000
20000
10000
0
CDKN1A
90
80
70
60
50
40
30
20
10
0
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
36.72 mb
36.71 mb
36.73 mb
T1
30
20
10
A1
2
1
T2
1000
500
T3
100
50
T4
80
40
T5
20
10
T6
30
20
10