Supplemental Material

Alternative splicing events as cancer drivers

Héctor Climente-González[1], Eduard Porta-Pardo[2], Adam Godzik[2], Eduardo Eyras[1,3]

[1]Universitat Pompeu Fabra (UPF), Barcelona, Spain

[2]Sanford Burnham Prebys Medical Discovery Institute, La Jolla, CA, 92037, USA

[3]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

Supplemental Data

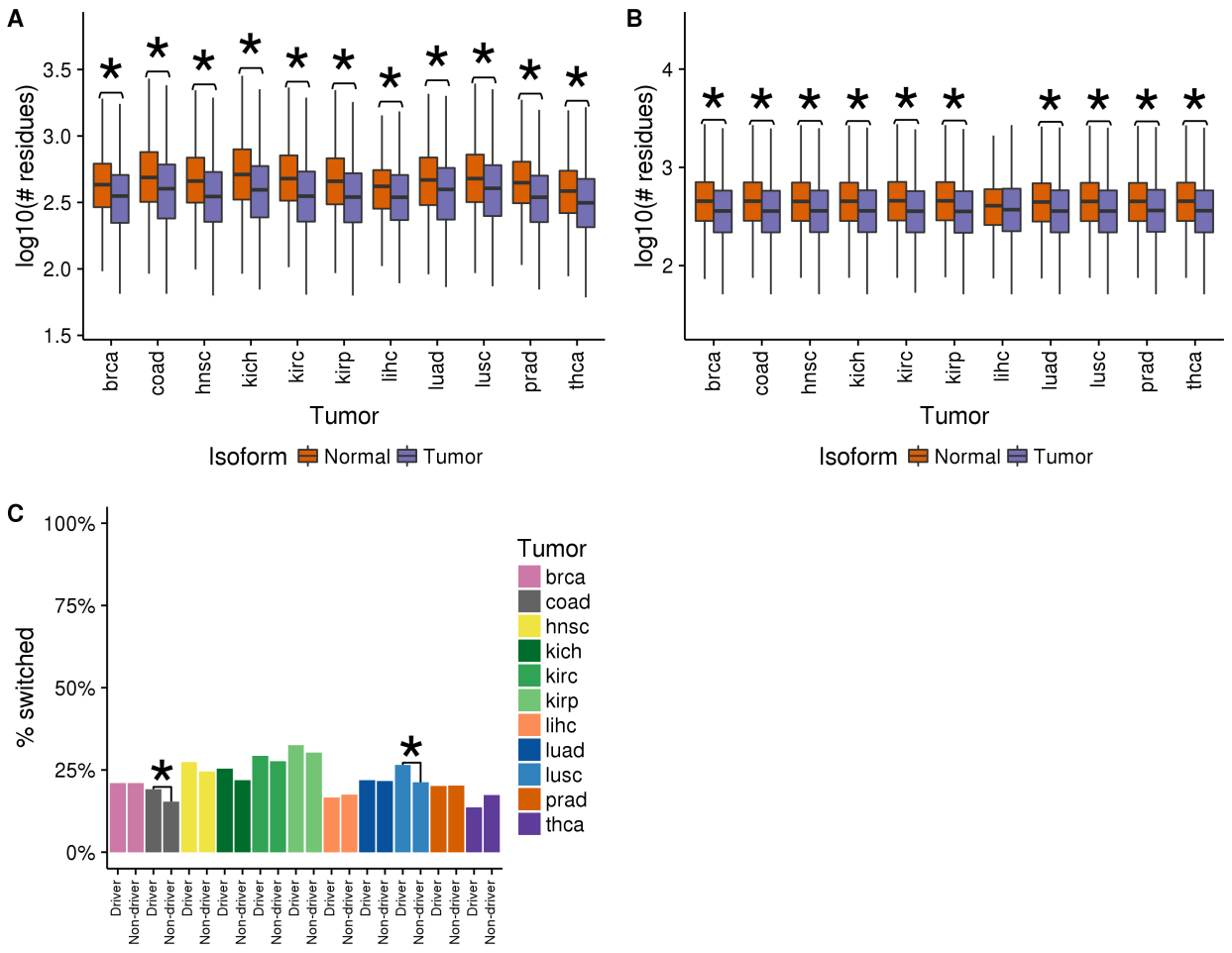

Figure S1. Properties of isoform switches in tumors. Related to Figure 1. **(A)** Distributions of the lengths of the tumor and normal protein isoforms in the calculated isoform switches. The *y*-axis indicates the number of residues in log10 scale. **(B)** Distributions of the lengths of the tumor and normal protein isoforms in the simulated alternative splicing switches. **(C)** For each tumor type, we give the proportions of genes with alternative transcript isoforms that present a switch, separated according to whether they are cancer drivers or not. Except for liver, prostate and thyroid tumors, in all cases there is a higher proportion of drivers with switches. We indicate with an asterisk those cases that are statistically significant.

Table S1. Properties of Isoform switches. Related to Figure 1. Provided as a text file with tab-separated values (.tsv). This table contains the list of identified isoform switches used for this analysis, including functional and nonfunctional ones, and AS-drivers. The table provides the following information:

Column number	Column label	Description
1	GeneId	Entrez gene id
2	Symbol	HGNC gene symbol
3	Normal_transcript	UCSC transcript id
4	Tumor_transcript	UCSC transcript id
5	Normal_protein	Swissprot ID (None if not known)
6	Tumor_protein	Swissprot ID (None if not known)
7	DriverAnnotation	“Driver” if it’s a driver, “d1” if it’s an interactor of a driver, and “Nothing” otherwise
8	IsFunctional	1 if it is functional as defined in the article, 0 otherwise
9	Driver	1 if it is a driver, 0 otherwise
10	Druggable	1 if it is a target of a known drug according to DGIdb (http://dgidb.genome.wustl.edu/)
11	CDS_Normal	1 if the normal transcript has an annotated CDS, 0 otherwise
12	CDS_Tumor	1 if the tumor transcript has an annotated CDS, 0 otherwise
13	CDS_change	1 if the CDS changes between the tumor and normal transcripts
14	UTR_change	1 if the 5’3 or 3’ UTRs change between the tumor and normal transcripts
15	Tumors	Tumor types in which the switch appears (brca, coad, etc…)
16	Number_samples	Number of samples in which the switch appears
17	Percentage_samples	Percentage of samples from the total studied across all tumor types in which the switch appears
18	Samples	IDs of samples in which the switch appears
19	Recurrence	1 if it is recurrent, 0 otherwise
20	PPI	1 if the switch affects a PPI in every tumor type where it appears; 0 otherwise. All PPIs affected by switches per tumor type are in Supp. File 3.
21	Affects_mutated_feature	1 if the switch leads to a gain or loss of a domain that is enriched in mutations in tumors, 0 otherwise
22	Pannegative	1 if the switch is mutually exclusive with 3 or cancer drivers and share a pathway with a cancer driver, 0 otherwise
23	AS_driver	1 if 19,20,21 or 22 is equal to 1, 0 otherwise
24	MS.pam	Samples with co-occurrence of switch and PAM in the same gene
25	M.pam	Samples with PAMs only
26	S.pam	Samples with Switches
27	N.pam	Rest of samples
28	p.pam.me	p-value of the mutual exclusion test
29	MS.mut	Samples with co-occurrence of switch and WGS mutations
30	M.mut	Samples with WGS mutations only
31	S.mut	Samples with Switches
32	N.mut	Rest of samples
33	p.mut.o	p-value of the co-occurrence of mutations and switches

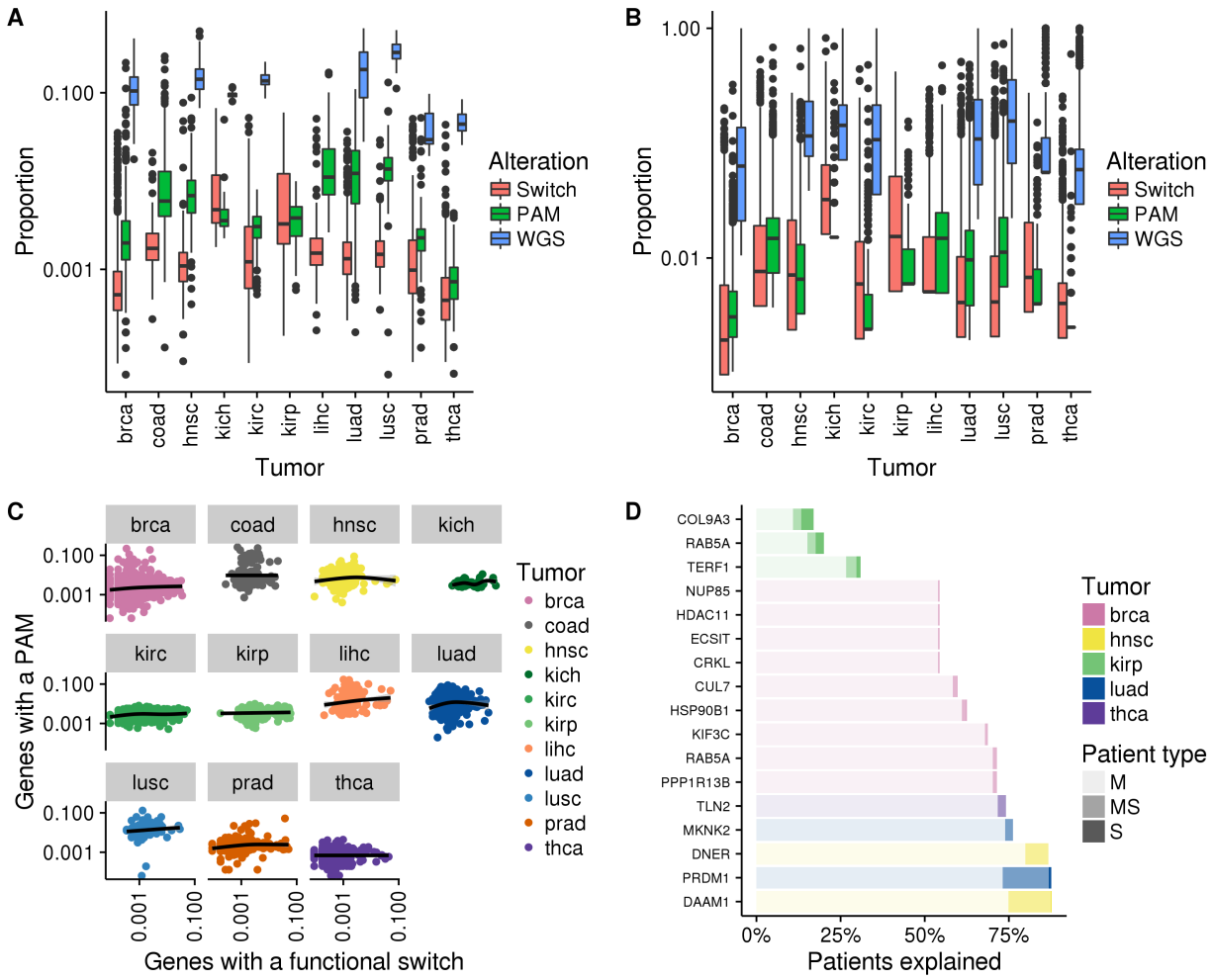

Figure S2. Properties of functional isoform switches in tumors. Related to Figure 2. **(A)** Proportion of genes in $\log_{10}$ scale (*y*-axis) with either of the three alterations: isoform switches (red), protein-affecting mutations (PAMs) (green), and any mutation type from whole genome sequencing (WGS) data (blue). **(B)** Proportion of samples (*y*-axis) with either of the three studied alterations: isoform switches (red), protein-affecting mutations (PAMs) (green), and any mutation type from whole genome sequencing (WGS) data (blue). **(C)** For each tumor type, each dot represents a sample according to the number of genes with a functional switch (*x*-axis) and the number of genes with protein-affecting mutations (PAMs) (*y*-axis). **(D)** Functional switches that potentially characterize pan-negative tumor samples. For each switch along the *y*-axis, we represent the proportion of patients from a given tumor type (*x*-axis) that harbor mutations in a tumor-specific mutational driver (M), have the switch (S), or have both (MS). The switches are ranked from the bottom of the *y*-axis according to the total number of patients explained. Only the top 30 cases are shown. Each case is color-coded according to tumor type. We considered as pan-negative those patients who do not harbor mutations in at least the top three mutated drivers in that tumor type.

Table S2. Mutation and domain gain/loss enrichments in protein domain families. Related to Figure 2. Provided as a text file with tab-separated values (.tsv). This table contains the information about the Protein domain families that are significantly enriched in mutations as well as gains or losses in isoform switches. The information provided for each domain family is the following:

Column number	Column label	Description
1	Pfam_id	PFAM ID for the domain family
2	Name	Name of the domain family
3	p_switch_gain	P-value for the gain-test
4	adjp_switch_gain	Adjusted P-value for the gain-test
5	p_switch_loss	P-value for the loss-test
6	adjp_switch_loss	Adjusted P-value for the loss-test
7	p_mutation	P-value for the mutation-test
8	adjp_mutation	Adjusted P-value for the mutation-test
9	Switches_where_gained	Number of switches where domain family is gained
10	Switches_where_lost	Number of switches where domain family is lost

Table S3. Mutual exclusion analysis between switches and cancer drivers. Related to Figure 2. Provided as a text file with tab-separated values (.tsv). This table contains the analysis of mutual exclusion between functional switches and mutational drivers in the same pathway:

Column number	Column label	Description
1	Tumor	Tumor type (brca, coad, etc…)
2	GeneId	Entrez gene ID
3	Symbol	HGNC gene symbol
4	Normal_transcript	UCSC transcript id
5	Tumor_transcript	UCSC transcript id
6	p_pannegative	P-value for the test for mutual exclusion (ME) with mutational drivers
7	Number_ME_drivers	Number of drivers with ME
8	ME_drivers	HGNC gene symbols for the ME drivers
9	Same_pathway_driver	Pathways shared with ME drivers
10	p_me_pathway_driver	P-value for the test for mutual exclusion (ME) with drivers in the same pathway

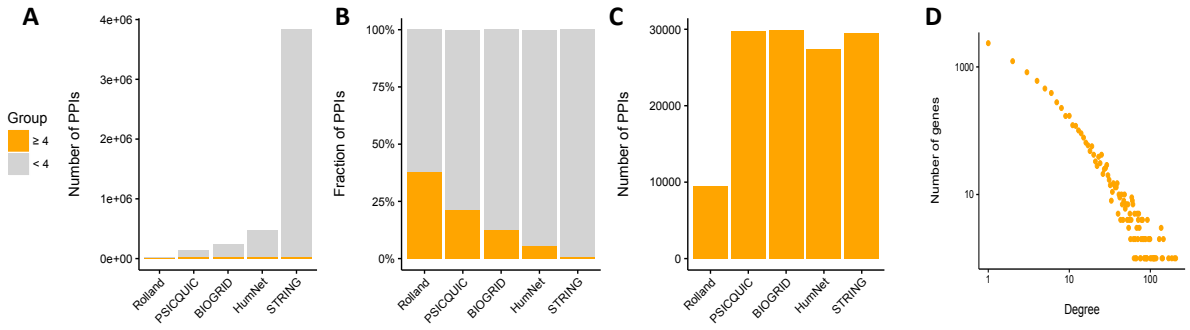

Figure S3. Protein-protein interaction network. Related to Figure 3**. (A)** Consensus protein–protein interaction (PPI) network. We used data from five different sources: PSICQUIC, BIOGRID, HumNet, STRING, and (Rolland et al., 2014). These networks vary in their size, connectivity, and origin, with PSICQUIC, BIOGRID, and Rolland being experimental networks and HumNet and STRING being functional networks. To build our consensus network, we used only those interactions that were defined in at least four different networks (shown in orange). **(B)** Fraction of each network included in the consensus network, with the data from (Rolland et al., 2014) having over 30% of its interactions and STRING less than 5%. **(C)** Number of interactions from each network included in the consensus network. **(D)** Degree distribution of the consensus network. For each number of PPI connections (*x*-axis), we give the number of genes with this degree (*y*-axis).

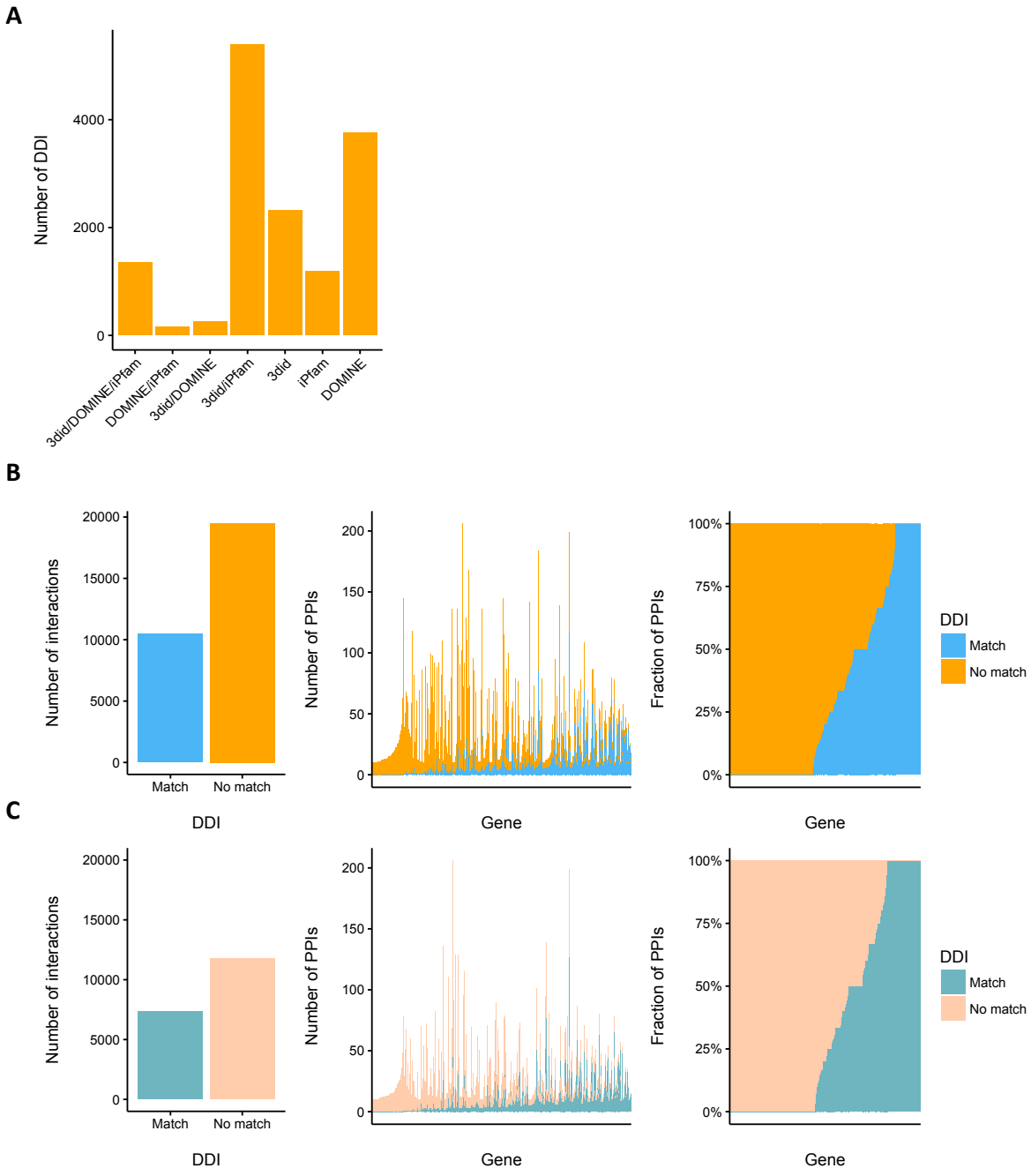

Figure S4. Protein-protein interactions assigned to functional isoform switches. Related to Figure 3**. (A)** Number of domain–domain interactions (DDIs) analyzed, separated by source: 3did, iPfam, DOMINE. **(B)** Protein interactions from the consensus network mapped to domain–domain interactions (DDIs). Left panel: We mapped to at least one DDI a total of 10,487 of the 29,991 interactions in the consensus network (35%). Middle panel: Absolute number of PPI interactions mapped (orange) or not mapped (blue) to a DDI in each gene (only genes with at least 10 PPIs are depicted). Genes are sorted according to the fraction of interactions that could

be mapped to DDIs. The picture shows no correlation between the degree of a gene and the fraction of interactions mapped. Right panel: Fraction of PPIs mapped to DDIs per gene. For 4,560 of the 8,142 genes in our network (56%), at least one interaction could be mapped. Genes are sorted according to the fraction of PPIs successfully mapped to DDIs. **(C)** These panels show the same as in (B) for genes with isoform switches.

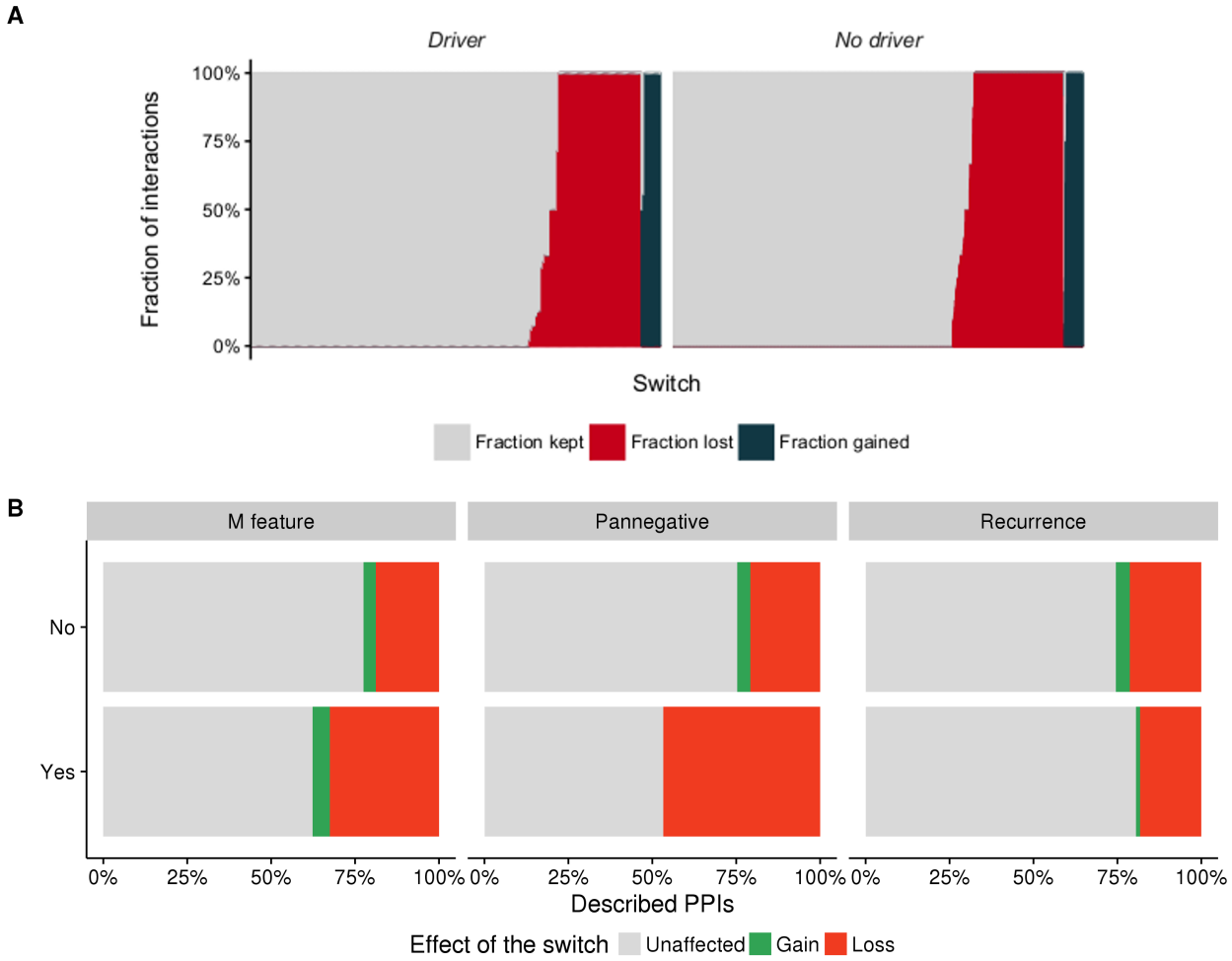

Figure S5. Properties of switches that affect protein-protein interactions. Related to Figure 3**. (A)** The figure depicts the fraction of interactions lost (red) or gained (black) for each switch with mapped domain–domain interactions (DDI) (*x*-axis). Interactions are separated according to whether they occur in drivers (left panel) or non-drivers (right panel). **(B)** Comparison of proportions of functional switches that affect protein-protein interactions. In the left panel, functional switches are divided according to whether they affect domains frequently mutated in cancer (M feature) (Yes) or not (No). In the middle panel, functional switches are divided according to whether the switch has mutual exclusion with tumor-specific drivers (pan-negative). In the right panel, functional switches are divided according to whether they are recurrent (Yes) or not (No). In each subset we plot the proportion of PPIs that are kept intact (gray), are lost (red), or are gained (green).

Table S4. Protein features and protein-protein interactions affected by isoform switches. Related to Figure 3. Provided as a text file with tab-separated values (.tsv). This table contains the proteins features and protein–protein interactions affected in each functional switch. The column descriptions are:

Column number	Column label	Description
1	Tumor	Tumor type (brca, coad, etc…)
2	GeneId	Entrez gene ID
3	Symbol	HGNC gene symbol
4	Normal_transcript	UCSC transcript id
5	Tumor_transcript	UCSC transcript id
6	Feature_type	Pfam, Prosite, IUPRED, ANCHOR
7	Pfam_id	ID for the protein feature if available
8	Name	Name of Feature if available, positions in protein for IUPRED and ANCHOR
9	Observation	Gained_in_tumor/Lost_in_tumor/No_change
10	Normal_isoform_order	Domain copy this corresponds to / total copies in normal isoform
11	Tumor_isoform_order	Domain copy this corresponds to / total copies in tumor isoform
12	GeneId_partner	Entrez ID of the protein-protein interaction partner
13	Symbol_partner	HGNC symbol of the protein-protein interaction partner
14	Transcript_partner	Transcripts identified as coding the interaction partner
15	Pfam_id_partner	PFAM ID for the domain mediating the interaction
16	Effect_on_interaction	Unaffected/Gain/Loss/NA(no interaction data)

Table S5. Pathways enriched in PPI-affecting switches. Related to Figure 3. Provided as a text file with tab-separated values (.tsv). This table contains the gene sets that are enriched in isoform switches that are predicted to affect protein-protein interactions. The enrichment tests is a Fisher's exact test based on the separations of switches being in the pathway or not, and affecting PPIs or not. We have tested Pathways, Complexes and gene sets-related to mRNA-metabolism. Only Pathways showed enrichment after multiple-test correction. The column descriptions are:

Column number	Column label	Description
1	Geneset_type	Pathway/Complex/mRNA_regulation
2	Geneset	Name of the gene set
3	Number_drivers	Number of drivers in the gene set.
4	p	Fisher's exact test p-value
5	adjp	p-value corrected for multiple testing
6	OR	Odds-ratio
7	eOR	Estimated odds-ration using with pseudocounts
8	Switched_genes	Genes in the gene set that have a PPI-affecting switch

Table S6. Gene modules with protein-protein interactions affected by isoform switches. Related to Figure 3. Provided as a text file with tab-separated values (.tsv). This table contains modules with high density of affect interactions: sets of genes that are connected in the network of protein-protein interactions and many of their interactions affected by the isoform switches and separately from other genes in the PPI network. We provide a test for assigning a complex or pathway based on the intersection of the complex/pathway to the module (see Experimental Procedures for details). The column descriptions are:

Column number	Column label	Description
1	Module	Module number
2	Module_components	Genes in the module (calculated from the network of protein-protein interactions affected by isoform switches)
3	Geneset	Name of complex/pathway compared to the module (NA if none was assigned)
4	Geneset_size	Number of genes in the complex/pathway (NA if none was assigned)
5	p	p-value from binomial test for the intersection of the gene set (Complex/Pathway) to the module
6	Intersection	Number of genes from the gene set that are in the module
7	Number_drivers	Number of cancer drivers in the module
8	padj	p-value corrected for multiple testing

Supplemental Experimental Procedures

Comparison with stromal and immune signatures

To determine whether the splicing changes observed merely reflect the cellular content of the samples, we measured the significant association with stroma and immune cell content using ESTIMATE (Yoshihara et al., 2013). For each switch we performed a Wilcoxon test to compare the ESTIMATE scores between patients with and without the switch. After correcting for multiple testing (Benjamini-Hochberg method), we found 819 and 240 exclusively associated (FDR < 0.05) with stromal and immune cell content, respectively; and 206 associated with both. These were eliminated from our analyses, as we could not be sure whether they are from the tumor cells.

Association between somatic mutations and isoform switches

Mutation data were downloaded from the TCGA data portal for all tumor types in the form of MAF files containing Level 2 somatic mutation calls from whole exome and Level 3 SNP array data. Additionally, we used somatic mutations from whole genome sequence (WGS) data for the 505 samples studied in (Fredriksson et al., 2014). For each transcript, its relative abundance or PSI was calculated as the fraction of the total gene expression (in TPM scale) explained by this transcript.

We compared the distributions of somatic mutations and functional switches across tumor samples. Using somatic coding and noncoding mutations from whole genome sequencing (WGS) data, the proportion of genes with multiple isoforms undergoing a functional switch per patient is on average about two orders of magnitude lower than the proportion of genes with mutations (Figure S2A). For protein-affecting mutations (PAMs), however, both distributions become more similar (Figure S2A). Likewise, the distribution of patients with switches is more similar to the distribution for PAMs than to the distribution for WGS mutations (Figure S2B).

We then tested the association of somatic mutations with functional switches by measuring the co-occurrence between switches and mutations as the proportion of patients with switches that also have mutations. This analysis showed higher proportions of functional switches and WGS mutations in cancer drivers compared to the rest of the genes (Wilcoxon test p-value < 0.0047). This higher co-occurrence is maintained when restricting to PAMs from WES data (p-value < 0.003). Nevertheless, only 16 (0.23%) functional switches show a significant association with somatic mutations (Fisher's exact test p-value < 0.05). Additionally, we found no correlation between the number of mutations and the number of switches per sample (Supplemental Figure 2C), and the majority of functional switches and PAMs in the same gene do not occur in the same sample. This suggests that the majority of functional switches may occur through trans-acting alterations (Sebestyén et al., 2016), such as the oncogenic splicing change in *RAC1* (Zhou et al., 2012), which is regulated by expression changes in hnRNP A1 and SR proteins (Gonçalves et al., 2009; Pelisch et al., 2012). On the other hand, genetic alterations affecting

splicing may involve small copy-number changes and micro-deletions in intronic regions that are hard to detect with WES and WGS data. As an example, a recurrent small deletion in TP73 was found recently in association to exon skipping in small cell lung cancer patients using WGS data (George et al., 2015). More targeted searches or deeper sequencing may be necessary to fully uncover such cases at genome-scale in other tumors.

Annotation of cancer gene drivers as oncogenes or tumor suppressors

We used the annotations provided by COSMIC (Forbes et al., 2015), by Vogelstein et al. (Vogelstein et al., 2013), and by the TSGene database (Zhao et al., 2015). Unlabeled cases were predicted with OncodriveROLE (Schroeder et al., 2014) using cutoffs 0.3 (loss-of-function class) and 0.7 (activating class).

Supplemental References

Forbes, S.A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S., Kok, C.Y., Jia, M., De, T., Teague, J.W., Stratton, M.R., McDermott, U., Campbell, P.J., 2015. COSMIC: Exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 43, D805–D811. doi:10.1093/nar/gku1075

Fredriksson, N.J., Ny, L., Nilsson, J.A., Larsson, E., 2014. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. Nat. Genet. 46, 1–7. doi:10.1038/ng.3141

George, J., Lim, J.S., Jang, S.J., Cun, Y., Ozretic, L., Kong, G., Leenders, F., Lu, X., Fernandez-Cuesta, L., Bosco, G., Muller, C., Dahmen, I., Jahchan, N.S., Park, K.-S., Yang, D., Karnezis, A.N., Vaka, D., Torres, A., Wang, M.S., Korbel, J.O., Menon, R., Chun, S.-M., Kim, D., Wilkerson, M., Hayes, N., Engelmann, D., Putzer, B., Bos, M., Michels, S., Vlasic, I., Seidel, D., Pinther, B., Schaub, P., Becker, C., Altmuller, J., Yokota, J., Kohno, T., Iwakawa, R., Tsuta, K., Noguchi, M., Muley, T., Hoffmann, H., Schnabel, P.A., Petersen, I., Chen, Y., Soltermann, A., Tischler, V., Choi, C., Kim, Y.-H., Massion, P.P., Zou, Y., Jovanovic, D., Kontic, M., Wright, G.M., Russell, P.A., Solomon, B., Koch, I., Lindner, M., Muscarella, L.A., la Torre, A., Field, J.K., Jakopovic, M., Knezevic, J., Castanos-Velez, E., Roz, L., Pastorino, U., Brustugun, O.-T., Lund-Iversen, M., Thunnissen, E., Kohler, J., Schuler, M., Botling, J., Sandelin, M., Sanchez-Cespedes, M., Salvesen, H.B., Achter, V., Lang, U., Bogus, M., Schneider, P.M., Zander, T., Ansen, S., Hallek, M., Wolf, J., Vingron, M., Yatabe, Y., Travis, W.D., Nurnberg, P., Reinhardt, C., Perner, S., Heukamp, L., Buttner, R., Haas, S.A., Brambilla, E., Peifer, M., Sage, J., Thomas, R.K., 2015. Comprehensive genomic profiles of small cell lung cancer. Nature 524, 47–53. doi:10.1038/nature14664

Gonçalves, V., Matos, P., Jordan, P., 2009. Antagonistic SR proteins regulate alternative splicing of tumor-related Rac1b downstream of the PI3-kinase and Wnt pathways. Hum. Mol. Genet. 18, 3696–3707. doi:10.1093/hmg/ddp317

Pelisch, F., Khauv, D., Risso, G., Stallings-Mann, M., Blaustein, M., Quadrana, L., Radisky, D.C., Srebrow, A., 2012. Involvement of hnRNP A1 in the matrix metalloprotease-3-dependent regulation of Rac1 pre-mRNA splicing. J. Cell. Biochem. 113, 2319–2329. doi:10.1002/jcb.24103

Rolland, T., Taşan, M., Charloteaux, B., Pevzner, S.J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., Kamburov, A., Ghiassian, S.D., Yang, X., Ghamsari, L., Balcha, D., Begg, B.E., Braun, P., Brehme, M., Broly, M.P., Carvunis, A.R., Convery-Zupan, D., Corominas, R., Coulombe-Huntington, J., Dann, E., Dreze, M., Dricot, A., Fan, C., Franzosa, E., Gebreab, F., Gutierrez, B.J., Hardy, M.F., Jin, M., Kang, S., Kiros, R., Lin, G.N., Luck, K., Macwilliams, A., Menche, J., Murray, R.R., Palagi, A., Poulin, M.M., Rambout, X., Rasla, J., Reichert, P., Romero, V., Ruyssinck, E., Sahalie, J.M., Scholz, A., Shah, A.A., Sharma, A., Shen, Y., Spirohn, K., Tam, S., Tejeda, A.O., Trigg, S.A., Twizere, J.C., Vega, K., Walsh, J., Cusick, M.E., Xia, Y., Barabási, A.L., Iakoucheva, L.M., Aloy, P., De Las Rivas, J., Tavernier, J., Calderwood, M.A., Hill, D.E., Hao, T., Roth, F.P., Vidal, M., 2014. A proteome-scale map of the human interactome network. Cell 159, 1212–1226. doi:10.1016/j.cell.2014.10.050

Schroeder, M.P., Rubio-Perez, C., Tamborero, D., Gonzalez-Perez, A., Lopez-Bigas, N., 2014. OncodriveROLE classifies cancer driver genes in loss of function and activating mode of action, in: Bioinformatics. doi:10.1093/bioinformatics/btu467

Sebestyén, E., Singh, B., Miñana, B., Pagès, A., Mateo, F., Pujana, M.A., Valcárcel, J., Eyras, E., 2016. Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. Genome Res. 26. doi:10.1101/gr.199935.115

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz Jr., L.A., Kinzler, K.W., 2013. Cancer Genome Landscapes. Science (80-.). 339, 1546–1558. doi:10.1126/science.1235122

Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., Treviño, V., Shen, H., Laird, P.W., Levine, D. a, Carter, S.L., Getz, G., Stemke-Hale, K., Mills, G.B., Verhaak, R.G.W., 2013. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat. Commun. 4, 2612. doi:10.1038/ncomms3612

Zhao, M., Kim, P., Mitra, R., Zhao, J., Zhao, Z., 2015. TSGene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. Nucleic Acids Res. 1–9. doi:10.1093/nar/gkv1268

Zhou, C., Licciulli, S., Avila, J.L., Cho, M., Troutman, S., Jiang, P., Kossenkov, a V, Showe, L.C., Liu, Q., Vachani, a, Albelda, S.M., Kissil, J.L., 2012. The Rac1 splice form Rac1b promotes K-ras-induced lung tumorigenesis. Oncogene 32, 903–909. doi:10.1038/onc.2012.99