RESOURCE

A single-cell strategy for the identification of intronic variants related to mis-splicing in pancreatic cancer

Emre Taylan Duman[1,#], Maren Sitte[1,#], Karly Conrads[2,3,5], Adi Makay[3,4], Fabian Ludewig[1], Philipp Ströbel[3,4], Volker Ellenrieder[2,3,6], Elisabeth Hessman[2,3,6], Argyris Papantonis[3,4,6,*], Gabriela Salinas[1,3,*]

[1] NGS-Core Unit for Integrative Genomics, Institute of Human Genetics, University Medical Center Göttingen, Germany
[2] Clinic of Gastroenterology, Gastrointestinal Oncology and Endocrinology, University Medical Center Göttingen, Germany
[3] Clinical Research Unit 5002 (KFO5002), University Medical Center Göttingen, Germany
[4] Institute of Pathology, University Medical Center Göttingen, Germany
[5] Institute of Medical Bioinformatics, University Medical Center Göttingen, Germany
[6] Comprehensive Cancer Center Lower Saxony (CCC-N), Göttingen, Germany
[#] These authors contributed equally to this work
* Correspondence to:
A.P., argyris.papantonis@med.uni-goettingen.de
G.S., gabriela.salinas@med.uni-goettingen.de

Summary

Most clinical diagnostic and genomic research setups focus almost exclusively on coding regions and essential splice sites, thereby overlooking other non-coding variants. As a result, intronic variants that can promote mis-splicing events across a range of diseases, including cancer, are yet to be systematically investigated. Such investigations would require both genomic and transcriptomic data, but there currently exist very few datasets that satisfy these requirements. We address this by developing a single-nucleus full-length RNA-sequencing approach that allows for the detection of potentially pathogenic intronic variants. We exemplify the potency of our approach by applying to pancreatic cancer tumor and tumor-derived specimens and linking intronic variants to splicing dysregulation. We specifically find that prominent intron retention and pseudo-exon activation events that are shared by the tumors and affect genes encoding key transcriptional regulators. Our work paves the way for the assessment and exploitation of intronic mutations as powerful prognostic markers and potential therapeutic targets in cancer.

Introduction

Introns in human genes can be extraordinarily large (up to 1 Mbp, with ~3,400 being longer than 50 kbp and ~1,200 longer than 100 kbp), and account for half of the non-coding human genome. As mutations in introns do not directly affect protein-coding sequences, they are usually overlooked [1–3]. As a result, little attention is paid to the importance of intronic-located splicing regulatory elements that control the fidelity of pre-mRNA splicing and transcription timing. This is surprising given that pathogenic variants causing abnormal splicing changes, typically by damaging existing splicing motifs or creating novel splicing motifs and may comprise 15–60% of all human disease variants [1–3].

Recent studies have reported significant numbers of intronic variants and deletions in protein-coding genes that are associated with under- or overexpression of the affected genes or of distant genes interacting in 3D, thus influencing their regulation in normal or pathogenic conditions[4–6]. It is also well established that introns contribute to the control of gene expression by including regulatory regions and non-coding (yet functional) genes or even directly by their extensive length[6]. A substantial number of pathogenic variants located deep within introns (i.e., >100 bp from an exon-intron boundary) were recently reported, suggesting that sequence analysis of full introns may help to identify causal mutations for many undiagnosed clinical cases[4–9]. Moreover, direct associations between intronic mutations and certain diseases have also been reported, albeit sporadically[7–9]. These results agree with the findings we present here. For instance, we identified genes that are simultaneously overexpressed in basal-like cells from pancreatic cancer (PDAC) tumors and rank as the most mutated transcripts, particularly when we consider intronic variants.

However, there exist many limitations when investigating intronic mis-splicing variants, the main ones being the lack of approaches capable of simultaneously interrogating genomic and transcriptional information, and the lack of guidelines designed for assessing intronic variants and their contribution to abnormal splicing changes in disease. We address these limitations by introducing a novel pipeline that utilizes full-length single-nuclei and bulk RNA sequencing strategy for the "deep" characterization of genetic variability within introns and of their effects on splicing and gene expression in PDAC. Indeed, cancer is one of those diseases, where alternative splicing is the basis for the identification of novel diagnostics, and therapeutic strategies for therapy (e.g., antisense oligonucleotides or small-molecule modulators of spliceosome[9–11]). This new approach therefore holds promise for both the elucidation of fundamental biological principles connected to splicing regulation, and the identification of therapeutic targets in human disease.

Results

Strategy for the identification of intronic mis-splicing variants affecting splicing in PDAC. One of the most critical post-transcriptional mechanisms reprogramming transcriptional output and proteomic diversity in cancer cells is the loss of splicing precision when removing introns from pre-mRNAs [12]. Consequently, the majority of mis- spliced variants are instead targeted for nuclear degradation or for nonsense-mediated mRNA decay (NMD) and thus, only few annotated alternative isoforms correspond to the precursors of the proteins mapped by large-scale proteomics studies[13].

To investigate this mechanism, we applied a new pipeline using a full-length snRNA-seq approach to three primary PDAC tumors (TM16, TM27 and TM56) and matched tumor-derived preclinical PDAC models (i.e., organoids and subcutaneous patient-derived Xenograft, PDXs). We then sought to determine pathogenic intronic variants causing abnormal splicing in these patient samples

(**Fig. 1a**). All samples were part of a unique patient cohort recruited for the Clinical Research Unit 5002 (https://gccc.umg.eu/en/cru-5002/).

For our snRNA-Seq experiments, we used the ICELL8 platform previously established in our group connecting genotype to phenotype in individual cells[15]. In contrast to the chemistry used by droplet-based platforms (i.e., in 3'-end approaches), ICELL8 is based on the SMART full-length chemistry allowing for the full read coverage of transcripts. Notably, when single nuclei are processed with this platform and chemistry, a strong enrichment in pre-mRNA is observed, including comprehensive coverage of introns and exons along these pre-mRNAs (**Fig. 1a and 2c**). Furthermore, utilization of full-length chemistry allowed strong detection of intergenic and intronic sequences, as well as of non-coding RNAs, especially long intergenic non-coding RNAs (lincRNAs)[14,15].

For the identification of intronic mis-splicing variants, we developed a variant-calling pipeline that combines i) snRNA-seq data determining the variant's position in the genome with ii) bulk RNA-seq data mainly capturing mRNAs and thus, charting splicing events in our samples (**Fig. 1b**).
The pipeline begins with a pseudo-bulk snRNA-seq variant calling analysis using the per-sample GATK method (**Fig. 1b (1-4)**), followed by an intersection of the individual findings to identify exact mutations present in all samples. The variants are then classified according to their genomic position taking into account intergenic, intronic and exonic regions using the maf file (Mutation Annotation File). By filtering the intronic variants, the genes were normalized and ranked according to intronic mutation load. The intronic variants are then forwarded to two parallel branches of the pipeline.
The first branch is set up for the investigation of these variants on the single cell level. Therefore, barcode-variant matrices containing the number of mutations for each gene (columns) and cell (rows) have been created to allow further transcriptional analysis related to the variants.
Simultaneously, the second branch of the pipeline connects genes that exhibit intronic mis-splicing to pathogenic splicing events through the investigation of partial or total intron retention (IR) or pseudo-exon activation (PEA) using the IRFinder algorithm (**Fig. 1b (5)**). Finally, the intron mutated genes from the snRNA data were merged with the intron retained genes from bulk RNA-Seq data. As an outcome, we highlight genes showing mis-splicing related intronic variants that contribute to malignancy and thus proposed as potential therapeutic targets in pancreatic cancer.

Identification of PDAC intronic variants using snRNA-seq. UMAP visualization of single nuclei clustering from the primary PDAC tumors, organoids and PDXs (**Fig. 2a**) was strongly influenced by the patient of origin of each sample. To evaluate the consistency in transcriptomics and intronic mis-splicing variants, we merged data from all PDAC models of each patient and regenerated separate UMAP clustering for TM16, TM27 and TM56 (**Fig. 2b**). Then, by applying our variant calling pipeline (**Fig. 1b**) to this snRNA-seq data, we observed that >95% of the mutations we identified were located in non-coding sequences (i.e., in intronic or intergenic regions) and only ~5% were found in protein-coding sequences. The most mutated genes at the level of introns included *RBFOX1*, *CSMD1*, *WWOX*, *CNTNAP2*, and *LRP1B* and were notably shared among PDAC tumors and models (**Fig. 2c**).

In order to identify *bona fide* intronic variants, we determined the Ti/Tv ratio in each primary tumor-, organoid- or PDX-derived dataset. Interestingly, primary tumor data showed higher precision and a more realistic Ti/Tv ratio (2.18) in comparison to organoid- and PDX-derived data (**Fig. 2d**). We therefore decided to focus on the three tumor-derived datasets for all ensuing analyses, which we complemented with bulk RNA-Seq data generated for 24 primary PDAC tumors from the CRU5002 cohort, as well as with public normal and tumor pancreas tissues data (accession ID: GSE211398). In the

end, variants shared between all three-tumor snRNA-seq datasets amounted to ~78,000 non-coding variants with 41,834 in introns, 25,449 in intergenic regions (IGR), 4,390 in 3' and 3,882 5' gene flanks, 610 in 3' UTRs, and 79 in 5' UTRs. Of these, 98.82% (n=76499) qualified as SNPs, 0.66% (n=510) as deletions, and 0.52% (n=404) as insertions (**Fig. 2e**). Regarding the genomic annotation, most (n=36,412) mapped in protein-coding genes or lncRNAs (n=10,926) and very few (n=121) in miRNAs (**Fig. 2c** and **Supplemental Table 1**).

Classification of intronic mis-splicing variants based on location. Next, to classify intronic variants we first performed a ranking based on the number of variants detected per gene normalized by the length and number of introns in each gene. Then, in an effort to link intronic variants to mis-splicing, we stratified mutations located 1–2, 3–20 or >20 nt away from the nearest exon-intron junction, which we classified as donor/acceptor sites (SS), branchpoint-proximal regions (BPs), polypyrimidine tracts (PPTs). All other variants qualified as "deep intronic" mutations (see **Table 1** and **Supplemental Table 2**)[5]. On the basis of these indexes, many genes showed a high number of deep intronic variants, and some of these were reported as tumor suppressors with a multitude of alternative splicing variants expressed in specific tissues or only in tumor cells, e.g., *LRP1B*, *CSMD1*, *WWOX*, *FHIT*, *MTUS2*[16–22] (**Fig. 3a**). The most prominent of these was *RBFOX1* and showed 208 "deep intronic" variants. The *RBFOX* family of RNA-binding proteins is well known to regulate alternative splicing (AS)[23,24]. Recently, *RBFOX2* was reported to modulate a metastatic AS signature in pancreatic cancer[25].

We next examined the fraction of intronic variants over all variants at each position (**Fig. 3b**). 99.6% of them were located deep in introns. Variants located in the proximal region are 0.32% of the total, with just 0.02% near SSs. By performing functional network analysis on the genes enriched for SS, proximal, and deep intronic regions, we identified enrichment of two main biological processes. The first, mRNA capping (GO:0006370), is most related to genes with variants in the SS and proximal region (e.g., *HNRNPH1, INTS10*, *TBPL1*, *POLR2L*, and *SNRNP35*). The second, regulation of transcription through histone methylation and H3K4-specific histone methyltransferase activity (GO:0051568), involved genes mainly carrying deep intronic variants (like *SMYD3*, *KMT2C*, *JMJD1C*, and *ASH2L*). Both GO terms are connected and the genes they contain are essential regulators gene expression and splicing[26]. Finally, visualization of the distribution of intronic variants in top mutated genes *WWOX*, *SMYD3* (**Fig. 3c**), *JMJD1C*, and *NAV* (**Supplemental Fig. 1**) exemplifies read coverage from snRNA-seq and bulk RNA-seq in primary PDAC tumors and how these can be superimposed to evaluate splicing events.

Transcriptional regulation relates to mis-splicing in basal-like tumor cells. We performed transcriptional subtyping of single tumor nuclei into the more aggressive/drug-resistant "basal-like" (BL) or the better prognosis-associated "classical" subtype (CLA) using a ranking markers method described previously [27] on snRNA-seq data. The distributions of each subtype in each tumor showed BL cells highly represented in primary tumors, and partially in matched PDXs. In contrast, CL cells predominated in our organoid models (**Fig. 4a**). Next, we performed differential expression analysis (DEG) between BL and CL tumor cells, where we identified a few hundred markers for both subtypes with absolute $\log_2$FC >0.5 and P_{adj} <0.05 (**Supplemental Table 3**) and visualized the most prominent ones in UMAP plots (**Fig. 4b**).

Tumor heterogeneity was assessed by evaluating cell types using snRNA-seq from all models at hand. As expected, primary tumors exhibited the highest heterogeneity as regards cell type composition while organoids mostly contained ductal-like cells (**Fig. 5a-c**). Strikingly, *RBFOX1*, *CSMD1*, and *CNTNAP2* that were our top most intronic mutated genes (**Figs 1b** and **4c**) appear to be simultaneously the genes most upregulated in BL cells (**Fig. 5a**). At the same time, the markers found in CL cells have already been described as biomarkers for pancreatic cancer: *MALAT1*, *NEAT1*, *CEACAM6* or *MUC1* [28–31]. For the

lncRNA *NEAT1*, two novel mutations are reported here for the first time (**Supplemental Table 1**). Taken together, CL cells and models are characterized by less abundance and relevance of intronic variants according to the output of our pipeline, thus suggesting that mis-splicing mechanisms are linked to the aggressiveness of BL tumors (**Fig. 5d**).

Contribution of intronic mis-splicing variants to pathogenic splicing and poor PDAC prognosis. In general, splicing events affect the regulation of genes (up- or downregulation) or generate new isoforms in normal and tumor tissues [32–34]. One example in our data, describing a non-pathogenic splice event for the *MACF1* gene, specifically the generation of a new isoform including two exon skipping events in pancreatic tissue, is given in **Supplemental Fig. 2**. However, our focus here is on potentially pathogenic splicing events for PDAC. Thus, we looked carefully into genes most enriched for intronic variants (**Table 1**) and simultaneously overexpressed in BL tumor cells (*RBFOX1*, *CSMD1*, and *CNTNAP2*). All three are long genes with a large number of (mostly small-, <100 nt, and micro-,<60 nt) exons, and with relatively small introns. For example, combinatorial splicing in a small portion of *CSMD1* (chr8: 3,187,000-3,202,000) showed an unusual exon skipping in the middle of the exon (chr8: 3,189,980-3,190,067) followed by two PEA events (chr8: 3,197,495-3,197,578 and 3,200,391-3,200,466; **Fig. 6a**). To evaluate the influence of intronic variants on these splicing events in *CSMD1*, we observed an enrichment of GGG motifs upstream of the PPT containing binding motifs for SRSF1 (**Fig. 6a**). Following motif analysis for SR-protein binding sites, we identified two intronic variants in our snRNA-seq data creating two *de novo* SRSF2 motifs [35]: i.e., C>G substitution: CA**C**GCT>CA**G**GCT; p.3,194,680 and CA**C**GAA>CA**G**GAA; p.3,196,860 (**Fig. 6a**). This suggests that intronic variants located deep in the intron and creating additional SR binding sites that may synergistically contribute to activate PEA events. To this day, few examples of PEA have been reported as caused by deep intronic mutations without directly changing a splice site sequence [36,37].

Similarly, looking into *HNRNPH1*, reported by our pipeline for carrying high number of intronic variants in proximal regions (**Table 1**), we could detect a combinatorial event involving four IR events (**Fig. 6b**). As HNRNP nucleoproteins are known to associate with pre-mRNAs in the nucleus and influence their processing and other aspects of mRNA metabolism and transport, our observations could be linked to their role in carcinogenesis, as reported recently for the alternative splicing of an oncogenic fusion transcript [38]. To identify genes that exhibit IR associated with intronic variants, we applied IRFinder on bulk RNA-seq data from 24 primary PDAC tumors recruited for CRU5002 and on public data (see **Fig. 1b**). A total of 2,489 shared and robust IR events were identified (listed in **Supplemental Table 4**); of those, 22 also carried multiple intronic variants according to our pipeline (green dots in **Fig. 7a**). Survival analysis performed using TCGA data that this group of 22 genes revealed their significant association with poor prognosis in PDAC (**Fig. 7b**). Six of these genes (*METTL3*, *HNRNPH1*, *INTS3*, *ELOB*, *EHMT1* and *KMT2C*) are linked to the GO terms of pre-mRNA capping (GO:0006370) and histone H3K4 modification (GO:0051568) discussed above.

Moreover, several transcripts encoding members of the Integrator complex (*INTS10*, *INTS3*, *INTS11*) were affected by multiple IR events (see *INTS3* examples in **Fig. 7c,d**). As the Integrator complex interacts with the C-terminal domain of RNA polymerase II to allow processing of U1 and U2 small nuclear RNAs, these splicing alterations could indirectly affect splicing in PDAC. Closer inspection of the two IR events in *INTS3* revealed an enrichment for mutations in snRNA-seq data that give rise to CCC sequences that follow a GGG motif. These give rise to partially overlapping recognition motifs for SRSF1 and -2 (**Fig. 7e**) and could function as splicing enhancers to compensate for weak PPTs tracts [39]. The high

intronic mutation load we uncovered using snRNA-seq suggests that the occurrence of such events at the single-cell level might be significantly more frequently than previously presumed [37].

Discussion

The current study proposes a new strategy for the investigation of intronic mis-splicing variants and their contribution to aberrant splicing promoting PDAC. In contrast to previous studies using whole genome sequencing and transcriptomic data, we present for a new full-length snRNA-seq approach allowing for the identification of intronic variants with high resolution. In fact, the use of full-length chemistry in the nuclei fraction results in an enrichment for pre-mRNA signal and therefore for the comprehensive read coverage of both introns and exons. By combining snRNA-seq with bulk RNA-seq data we could then determine pathogenic splicing events that are related to these intronic variants.

We developed a comprehensive pipeline to take advantage of integrating multi-omics information e.g. variant calling, transcriptomics and splicing within a cell and thus offering a more holistic view of the underlying molecular mechanisms in complex diseases. At the same time we address the challenges applying a calling variant pipeline on snRNA-Seq data handling with low-covered regions, non-uniform read distribution and thus, increased false positive findings. Accordingly, we performed a deep coverage sequencing approach for both data sets, namely 300 to 400K median reads per nuclei (snRNA-Seq) and 70 to 100 Mio reads per sample, PE 300 cycles (bulk RNA-Seq). Further, we followed the current GATK recommendation for RNA-seq data performing variant calling from individual samples (per-sample GATK method) [40]. In addition, only intersected variants found in all primary PDAC were assessed reducing false positive findings as indicated in supplementary Table 1.

Our approach delivers extensive information on the landscape of primary transcripts especially near pathogenic splicing events. Thus, we can begin to decipher the complexity of RNA sequences acting as repressors or activators of splicing in the context of PDAC. Recent high throughput characterization of exon splicing enhancer and silencer (ESE and ESS) motifs has indicated that pseudo-exons can be discriminated from genuine exons on the basis of their low ESE and high ESS contents [41,42]. Moreover, these motifs can be found both in exons and in introns, and function via the recruitment of sequence-specific RNA-binding proteins that can dictate splicing choices [41–43]. Our analyses suggest that splicing dysregulation in PDAC can be linked to non-canonical sequence signals, i.e., to intronic variants affecting RNA motifs located deep in introns. For instance, we found several motifs for SR RNA-binding proteins (e.g., SRSF1 and SRSF2), enriched for GGG and CCC, distributed in the SS-proximal regions of introns and, in some cases, exhibiting extensive overlap (**Fig. 7**) [35,44]. Although RNA-binding proteins like SR proteins bind to RNA with high sequence specificity, it is difficult to obtain well-defined consensus motifs for each of them [35,45]. However, cooperation and competition between the SRSF1 and SRSF2 proteins have been reported in the regulation of alternative splicing events, which are related to synergistic and compensatory interactions with target RNA [35]. Still, the exact mechanism by which these variants affect the splicing machinery and its downstream choices remains to be dissected, and very few examples of deep-intronic disease-causing mutations have been described to date [36,37].

The complex nature of splicing events and their regulation necessitates the development and implementation of new approaches to uncover nuanced relationships between mutations and splicing

choices. We described a pseudo-bulk variant-calling pipeline exploiting dense and full-length coverage of snRNA-seq datasets to detect putatively pathogenic deep intron variants and link them to mis-splicing events occurring in PDAC tumors, while also gauging cell heterogeneity. As our method generates a multidimensional output that typically serves as the foundation for machine learning models (ranging from Support Vector Machines to Convolutional Neural Networks), we envisage a near-future combination that would vastly improve our understanding of the combinatorial code regulating splicing choices in the context of cancer.

References

1. Shiraishi, Y. *et al.* Systematic identification of intron retention associated variants from massive publicly available transcriptome sequencing data. *Nature Communications 2022 13:1* **13**, 1–13 (2022).

2. Zhang, P. *et al.* Genome-wide detection of human variants that disrupt intronic branchpoints. *Proc Natl Acad Sci U S A* **119**, e2211194119–e2211194119 (2022).

3. Park, E., Pan, Z., Zhang, Z., Lin, L. & Xing, Y. The Expanding Landscape of Alternative Splicing Variation in Human Populations. *Am J Hum Genet* **102**, 11–26 (2018).

4. Le Hir, H., Nott, A. & Moore, M. J. How introns influence and enhance eukaryotic gene expression. *Trends Biochem Sci* **28**, 215–220 (2003).

5. Jung, H., Lee, K. S. & Choi, J. K. Comprehensive characterisation of intronic mis-splicing mutations in human cancers. *Oncogene 2021 40:7* **40**, 1347–1361 (2021).

6. Daguenet, E., Dujardin, G. & Valcárcel, J. The pathogenicity of splicing defects: mechanistic insights into pre-mRNA processing inform novel therapeutic approaches. *EMBO Rep* **16**, 1640 (2015).

7. Scotti, M. M. & Swanson, M. S. RNA mis-splicing in disease. *Nat Rev Genet* **17**, 19–32 (2016).

8. Qian, X. *et al.* Identification of Deep-Intronic Splice Mutations in a Large Cohort of Patients With Inherited Retinal Diseases. *Front Genet* **12**, 276 (2021).

9. Sahin, I., George, A. & Seyhan, A. A. Therapeutic Targeting of Alternative RNA Splicing in Gastrointestinal Malignancies and Other Cancers. *Int J Mol Sci* **22**, (2021).

10. Raguraman, P., Balachandran, A. A., Chen, S., Diermeier, S. D. & Veedu, R. N. Antisense Oligonucleotide-Mediated Splice Switching: Potential Therapeutic Approach for Cancer Mitigation. *Cancers (Basel)* **13**, (2021).

11. Peng, Q. *et al.* Impacts and mechanisms of alternative mRNA splicing in cancer metabolism, immune response, and therapeutics. *Molecular Therapy* **30**, 1018–1035 (2022).

12. Bonnal, S. C., López-Oreja, I. & Valcárcel, J. Roles and mechanisms of alternative splicing in cancer - implications for care. *Nat Rev Clin Oncol* **17**, 457–474 (2020).

13. Tress, M. L., Abascal, F. & Valencia, A. Alternative Splicing May Not Be the Key to Proteome Complexity. *Trends Biochem Sci* **42**, 98–110 (2017).

14. Wang, X., He, Y., Zhang, Q., Ren, X. & Zhang, Z. Direct Comparative Analyses of 10X Genomics Chromium and Smart-seq2. *Genomics Proteomics Bioinformatics* **19**, 253–266 (2021).

15. Shomroni, O. *et al.* A novel single-cell RNA-sequencing approach and its applicability connecting genotype to phenotype in ageing disease. *Sci Rep* **12**, 4091–4091 (2022).

16. Kuo, H. C. *et al.* A Comparative Proteomic Analysis of Erinacine A's Inhibition of Gastric Cancer Cell Viability and Invasiveness. *Cellular Physiology and Biochemistry* **43**, 195–208 (2017).

17. Huang, T. *et al.* Csmd1 mutations are associated with increased mutational burden, favorable prognosis, and anti-tumor immunity in gastric cancer. *Genes (Basel)* **12**, (2021).

18. Chen, H. *et al.* Association of LRP1B mutation with tumor mutation burden and outcomes in melanoma and non-small cell lung cancer patients treated with immune check-point blockades. *Front Immunol* **10**, (2019).

19. Datta, N., Chakraborty, S., Basu, M. & Ghosh, M. K. Tumor Suppressors Having Oncogenic Functions: The Double Agents. *Cells* **10**, 1–26 (2021).

20. Del Mare, S., Salah, Z. & Aqeilan, R. I. WWOX: its genomics, partners, and functions. *J Cell Biochem* **108**, 737–745 (2009).

21. Salah, Z., Aqeilan, R. & Huebner, K. WWOX gene and gene product: tumor suppression through specific protein interactions. *Future Oncol* **6**, 249 (2010).

22. Iliopoulos, D. *et al.* Roles of FHIT and WWOX fragile genes in cancer. *Cancer Lett* **232**, 27–36 (2006).

23. Sun, S., Zhang, Z., Fregoso, O. & Krainer, A. R. Mechanisms of activation and repression by the alternative splicing factors RBFOX1/2. *RNA* **18**, 274–283 (2012).

24. Lin, F. *et al.* Identification of inflammatory response and alternative splicing in acute kidney injury and experimental verification of the involvement of RNA-binding protein RBFOX1 in this disease. *Int J Mol Med* **49**, (2022).

25. Jbara, A. *et al.* RBFOX2 modulates a metastatic signature of alternative splicing in pancreatic cancer. *Nature* (2023) doi:10.1038/S41586-023-05820-3.

26. Cowling, V. H. Regulation of mRNA cap methylation. *Biochem J* **425**, 295–302 (2009).

27. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nature Methods 2017 14:11* **14**, 1083–1086 (2017).

28. Xiong, D. D. *et al.* The clinical value of lncRNA NEAT1 in digestive system malignancies: A comprehensive investigation based on 57 microarray and RNA-seq datasets. *Oncotarget* **8**, 17665 (2017).

29. Duxbury, M. S. *et al.* CEACAM6 is a novel biomarker in pancreatic adenocarcinoma and PanIN lesions. *Ann Surg* **241**, 491–496 (2005).

30. Striefler, J. K. *et al.* Mucin-1 Protein Is a Prognostic Marker for Pancreatic Ductal Adenocarcinoma: Results From the CONKO-001 Study. *Front Oncol* **11**, 1 (2021).

31. Kumar, S. & Mishra, S. MALAT1 as master regulator of biomarkers predictive of pan-cancer multi-drug resistance in the context of recalcitrant NRAS signaling pathway identified using systems-oriented approach. *Sci Rep* **12**, (2022).

32. Wong, J. J. L., Au, A. Y. M., Ritchie, W. & Rasko, J. E. J. Intron retention in mRNA: No longer nonsense: Known and putative roles of intron retention in normal and disease biology. *Bioessays* **38**, 41–49 (2016).

33. Jung, H. *et al.* Intron retention is a widespread mechanism of tumor-suppressor inactivation. *Nature Genetics 2015 47:11* **47**, 1242–1248 (2015).

34. Ge, Y. & Porse, B. T. The functional consequences of intron retention: alternative splicing coupled to NMD as a regulator of gene expression. *Bioessays* **36**, 236–243 (2014).

35. Pandit, S. *et al.* Genome-wide analysis reveals SR protein cooperation and competition in regulated splicing. *Mol Cell* **50**, 223–235 (2013).

36. Davis, R. L., Homer, V. M., George, P. M. & Brennan, S. O. A deep intronic mutation in FGB creates a consensus exonic splicing enhancer motif that results in afibrinogenemia caused by aberrant mRNA splicing, which can be corrected in vitro with antisense oligonucleotide treatment. *Hum Mutat* **30**, 221–227 (2009).

37. Homolova, K. *et al.* The deep intronic c.903+469T>C mutation in the MTRR gene creates an SF2/ASF binding exonic splicing enhancer, which leads to pseudoexon activation and causes the cblE type of homocystinuria. *Hum Mutat* **31**, 437–444 (2010).

38. Vo, T. *et al.* HNRNPH1 destabilizes the G-quadruplex structures formed by G-rich RNA sequences that regulate the alternative splicing of an oncogenic fusion transcript. *Nucleic Acids Res* **50**, 6474–6496 (2022).

39. Murray, J. I., Voelker, R. B., Henscheid, K. L., Warf, M. B. & Berglund, J. A. Identification of motifs that function in the splicing of non-canonical introns. *Genome Biol* **9**, 1–17 (2008).

40. Brouard, JS., Schenkel, F., Marete, A. *et al.* The GATK joint genotyping workflow is appropriate for calling variants in RNA-seq experiments. *J Animal Sci Biotechnol* **10**, 44 (2019).

41. Grellscheid, S.-N. & Smith, C. W. J. An Apparent Pseudo-Exon Acts both as an Alternative Exon That Leads to Nonsense-Mediated Decay and as a Zero-Length Exon. *Mol Cell Biol* **26**, 2237 (2006).

42. Han, K., Yeo, G., An, P., Burge, C. B. & Grabowski, P. J. A Combinatorial Code for Splicing Silencing: UAGG and GGGG Motifs. *PLoS Biol* **3**, e158 (2005).

43. Singh, R. RNA–Protein Interactions That Regulate Pre-mRNA Splicing. *Gene Expr* **10**, 79 (2002).

44. Rahman, M. A., Lin, K. T., Bradley, R. K., Abdel-Wahab, O. & Krainer, A. R. Recurrent SRSF2 mutations in MDS affect both splicing and NMD. *Genes Dev* **34**, 413–427 (2020).

45. Jankowsky, E. & Harris, M. E. Specificity and nonspecificity in RNA–protein interactions. *Nature Reviews Molecular Cell Biology 2015 16:9* **16**, 533–544 (2015).

46. Amezquita, R. A. *et al.* Orchestrating single-cell analysis with Bioconductor. *Nature Methods 2019 17:2* **17**, 137–145 (2019).

47. Chan-Seng-Yue, M. *et al.* Transcription phenotypes of pancreatic cancer are driven by genomic events during tumor evolution. *Nature Genetics 2020 52:2* **52**, 231–240 (2020).

48. Segerstolpe, Å. *et al.* Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metab* **24**, 593–607 (2016).

49. Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature Immunology 2019 20:2* **20**, 163–172 (2019).

50. Van der Auwera, G. A. *et al.* From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Curr Protoc Bioinformatics* **43**, 11.10.1-11.10.33 (2013).

51. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 1–14 (2016).

52. Peterson, H., Kolberg, L., Raudvere, U., Kuzmin, I. & Vilo, J. gprofiler2 -- an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler. *F1000Research 2020 9:709* **9**, 709 (2020).

53. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* **6**, (2013).

54. Middleton, R. *et al.* IRFinder: Assessing the impact of intron retention on mammalian gene expression. *Genome Biol* **18**, 1–11 (2017).

Acknowledgements. We are grateful for excellent technical support provided by Jacqueline Fink and Waltraut Kopp, and to Hanibal Bohnenberger for technical advice and for providing patient-derived samples. We also thank Jeniffer Appelhans and Mark Bösherz for the generation of organoids derived from primary PDAC tumors.

Funding. This work was supported by the Deutsche Forschungsgemeinschaft via KFO5002 (awarded to G.S., P.S., A.P., V.E. and E.H.).

Author Contributions. G.S., A.P., and E.H. conceived the project and experiment design. V.E and P.S supported the experimental design and contributed to experimental data G.S. led the method development. F.L. led the experimental data production. A.M. contributed to experimental data, E.T.D. and M.S. led and performed the data analysis. K.C supported data analysis and A.P. and G.S. wrote the manuscript. All authors have read and agreed to the submitted version of the manuscript.

Competing interests. The authors declare no competing interests.

Institutional review board statement. The utilization and characterization of human PDAC data and samples within the CRU5002 have been approved by the ethical review board of the UMG (11/5/17).

Informed consent statement. Informed consent was obtained from all subjects involved in the study.
.

Methods

Patients' sample information. Utilization and characterization of human PDAC data and samples within the CRU5002 has been approved by the ethical review board of the UMG (11/5/17). Sequencing studies and the generation of organoids and PDX models have been performed using tumor tissue from CRU5002 PDAC patients with progressed disease upon histological PDAC confirmation.

Generation of PDX models. For the generation of PDX models, bulk tumor tissue was subcutaneously transplanted into SHO-*prkdc*^{scid}*Hr*^{hr} mice. Engrafted subcutaneous tumors were passaged in mice for three generations prior to snRNA-sequencing.

Generation of organoids models. Tumor tissue was minced and digested in Dulbecco's modified Eagle's medium containing 5 mg/ml of collagenase XI (Sigma-Aldrich, C9407), DNAse final concentration 10µg/ml (SIGMA D5025-150KU) and Y-27632 final concentration 10,5µM (Adooq Bioscience, 129830-38-2) and incubated at 37°C for 45 min. The material was further embedded in Matrigel (Corning, New York, USA; Cat#356231) and cultured in human pancreatic cancer complete medium (Wnt3a, R-spondin1, Gastrin, hEGF, A 83-01, hFGF-10, mNoggin, Primocin, N-acetylcystein, Nicotinamide, B27 supplement and Y-27632). For passage, the Matrigel-containing organoid was digested by TrypLE™ Express (Thermo Fisher, 12605-028) with DNAse and Y-27632 as described above for 15 min. The sample was centrifuged at 500g for 5 min, and the precipitated cells were embedded in GFR Matrigel and cultured in human pancreas organoid complete feeding medium.

Nuclei extraction from primary PDAC and PDAC models. Nuclei isolation of PDAC was performed according to the "Nuclei Isolation from Cell Suspensions & Tissues for Single Cell RNA sequencing", Document Number CG000124 Rev E, 10x Genomics, (2021, June 30). For organoids, minor modifications were performed. The cells were spun down at 500 x g, lysis took place for 15 minutes and only two washing steps were performed. Following the second wash, the nuclei pellet was resuspended in 1 ml wash buffer in preparation for the ICELL8 protocol.

Full-length single-cell RNA-seq using the ICELL8. The Takara ICELL8 5,184 nano-well chip was used with the full-length SMART-Seq ICELL8 Reagent Kit. Nuclei suspensions were fluorescent-labelled with Hoechst 33,342 for 15 min prior to their dispensing into the Takara ICELL8 5,184 nano-well chips. Cell Select Software (Takara Bio) was used to visualize and select wells containing single nuclei. A total of nine 5,184 nano-wells chips were used for all samples and a total of 11,084 nuclei were processed for data analysis. Specifically, after quality control, 3,416 nuclei were used for primary PDAC, 3,037 for PDX, and 4,631 for organoids respectively. cDNA synthesis and library preparation were done according to description in previous study [12]. Libraries were sequenced on the HiSeq 4000 (Illumina) to obtain on average~0.3 Mio reads per nuclei (SE; 50 bp).

Bulk-RNA-Seq from primary PDAC samples. Total RNA was extracted from FFPE tumor patient samples using the ReliaPrep™ FFPE Total RNA Miniprep System (Promega). RNA Integrity was determined using the Fragment Analyzer. Because of low RNA integrity (sizing from 50 to 140 bp), we performed a modified TruSeq Stranded Total RNA Library Prep Human/Mouse/Rat (Cat. N°20020596) starting with 200 ng of total RNA. The modifications include a) ignoring fragmentation step, b) ligation optimization by adjusting adapters concentration during library preparation, c) increasing PCR cycles (15 in total) and eliminating primer dimers prior to sequencing (Agencourt AMPure XP magnetic beads, Beckman Coulter). Primary PDAC were sequenced on the NovaSeq6000; S4 flow cell PE 300 cycles generating a data set of 50 to 400 Mio reads per sample.

Bulk-RNA-Seq from organoids and PDXs. RNA libraries were prepared starting with 300 ng of total RNA using a non-stranded mRNA Seq (TruSeq RNA Library Preparation Cat. RS-122–2001) from Illumina

according to the manufacturer's recommendations. Libraries were sequenced on the Illumina HiSeq 4000 (SE; 1×50 bp; 30–35 Mio reads/sample).

Pre-processing of single-nuclei RNA-seq data. Raw sequencing files were processed as described in [12]. Briefly, Cogent NGS analysis pipeline (CogentAP) from Takara Bio (v1.0) was applied for de-multiplexing and creating the gene expression matrices from each FASTQ file. Reads were aligned against the human genome GRCh38 v107 (https://www.ensembl.org/Homo_sapiens/Info/Index). Raw matrices underwent quality-control (QC) filtering for cells and genes as outlined in [12] also including intron regions of genes. Gene matrices for cells that passed the quality control were used as input for the SingleCellExperiment R package (v3.0)[46] to generate SingleCellExperiment objects for the subsequent downstream analysis.

Identification of disease subtypes and cell types. To identify the tumor subtypes and cell types of which the single cells are composed, two different methods were employed. First, we used the marker-based method AUCell [27] to identify the PDAC subtypes classical and basal-like tumor. This analysis was based on established marker genes [47]. AUCell ranks the genes in each cell by decreasing expression value, and marks cells according to their most expressed marker genes. Secondly, we performed cell type annotation for more refined subpopulations to address the heterogeneity of the tumor and its matched models. We utilized the reference data set as provided by the SeuratData R package (panc8.SeuratData) [48] and the prediction function as implemented in the R package SingleR [49]. Downstream analyses performing the UMAP algorithm were done as implemented in CogentDS (v1.0) for dimensionality reduction and data visualization. To determine which genes were differentially expressed between tumor subtypes and cell types in a particular patient, Wilcoxon Rank Sum and Signed Rank Test was used, together with *p* values adjustment with Benjamini–Hochberg method.

Pseudo-variant calling in snRNA and bulk RNA-seq data. Bam files resulting from CogentDS were used as input for the pseudo bulk variant calling using GATK best practices pipeline for RNAseq variant calling [50]. Consistent with the recommendations of GATK, duplicates were removed with Picard MarkDuplicates, and read groups were added with Picard AddOrReplaceReadGroups. Subsequently, Cigar reads were split into exon segment and hard-clip any sequences overhanging into the intronic regions with SplitNCigarReads. Variant calling was performed by HaplotypeCaller and all variants were then hard filtered by the following criteria: FS>30 and QD<2. Patient specific vcf files were intersected or merged using bcftools-isec to collect shared or unique mutations from the samples. Resulting variants were converted into maf (mutation annotation file) using Ensembl VEP [51] annotation tool for the identification of the intronic, intergenic, and splice junction mutations (Fig.3-A). Finally, mutations were grouped into three categories: SS, Proximal, and Deep by using Ensembl GTF file version 107. Total number of SS and Prox mutations were subtracted from total number of mutations in maf files to calculate the number of deep intronic variants. Negative values indicate that the deep intronic mutation affects multiple transcripts of the gene. Pre-Ranked enrichment analysis was performed by using mutation lists from three locational groups. Mutation numbers from SS and Proximal regions were normalized by using total intron number of each gene. Deep mutation numbers were normalized by per unit length (kb). Normalized values are ranked and used in enrichment analysis by using R package gProfileR[52]. Overall survival was calculated using Kaplan-Meier analysis based on TCGA-PAAD cohort. The result is shown as Kaplan-Meier plot with p value from log rank test generated by the cBioPortal[53].

Cell specific variant calling. To perform single cell specific variant calling, bam files were separated into chip- and sample-specific bam files. Separated bam files were used for the variant calling pipeline which

was applied to individual bam file as before, without the base quality recalibration step (BQSR) to be able to get more variants possible per cell. Quality filtering was applied with the same quality filters as mentioned in pseudo-bulk RNA variant calling in the methods section. VCF files for each barcode were used to collect barcode-specific mutations and visualize as a mutation distribution plot using CogentDS UMAP functions.

Identification of intron retention events from RNA-Seq data. Intron retention analysis was performed using IRFinder 1.3.0 [54]. IRFinder calculates IR-ratios to measure IR level reflecting the proportion of intron retaining transcripts. To compare PDAC samples to normal tissue, we downloaded fastq files from nine PDAC and nine healthy human pancreas samples from NBCI GEO database (accession ID: GSE211398). These were pre-processed using the same methods as described in Bulk-RNA-Seq from primary PDAC. Tumor samples from GSE211398 dataset were then analyzed using the IRFinder pipeline to identify IR events. Intronic depth for all samples normalized by CPM due to the library size variations between the public data and CRU samples. Due to the different number of samples between the two dataset, weighted means of IR scores were calculated to balance the variable sample sizes by using 0.1 threshold among 80% of all samples. Normalized intron depths were also used as further filtering to reduce the number of identified possible false negative discoveries.

Data availability. All sequencing data has been deposited in the Gene Expression Omnibus, Bulk RNA-Sequencing GSE228844 and snRNA-Sequencing GSE229007. Any other relevant data is available from the authors upon reasonable request.

Code availability. Scripts used to analyze the data and generate the figures and tables in this paper are available on GitHub (https://github.com/UKHG-NIG/VarSpliceDetect).

Supplemental Information. The following supporting information accompanies this manuscript.

Supplemental Fig. 1: Intronic variant distribution and location for *NAV2* and *JMJCD1* assessed using pseudo-bulk snRNA-seq data.

Supplemental Fig. 2: Visualization of a non-pathogenic exon skipping event in *MACF1*.

Supplemental Fig. 3: Partial and total IR in *KMT2C* and *JMJD1C*.

Supplemental Table 1: Global intronic variants detected from the pseudo-bulk snRNA data.

Supplemental Table 2: Number of mutations based on locational classification (Deep-SS-Prox) with normalized numbers by intron length and number.

Supplemental Table 3: List of DEGs from classical and basal-like cells obtained from snRNA-seq data of primary PDAC tumors (n=3) and their corresponding PDXs (n=3) and organoids (n=3).

Supplemental Table 4: Genes with shared intronic mutations and IR events based on median CPM-normalized reads depth and median IR event occurrence.

Figure legends

Fig. 1: Strategy for the identification of intronic variants related to splicing in PDAC.

a, snRNA-seq using the ICELL8 platform and the SMART full-length chemistry allowing for full transcript coverage was applied to PDAC primary tumors (TM16, TM27 and TM56) and matched organoids and PDX lines. A total of nine 5184-nanowell chips were performed (three for primary PDAC, three for organoids and three for PDX respectively and visualized in the UMAP. Finally data from all nine nanowell chips were merged. Bulk RNA sequencing was additional performed for each of the samples.

b, Variant calling pipeline combining snRNA-seq and bulk RNA-seq data determining intronic variant positions in the genome and intron retention events. 1) Gold standard RNA-seq variant calling pipeline pre-processing; 2) sample-level variant calling; 3) combination strategies for variant discovery (Shared-Combined); 4) Further filtering for RNA-editing variant removal; and 5) determination of shared intron retention events with weighted IR-ratio calculation between public and own PDAC samples.

Fig. 2: Identification of intronic variants using snRNA-seq data

a, UMAP plot for three primary PDAC tumors and corresponding models (TM16, TM27 and TM56).

b, UMAP plot showing clustering by patients (primary PDAC and models) forTM16, TM27 and TM56.

c, Pie charts of the three PDACs (tumor, organoids and PDX) showing percentage of exonic and intronic mutations (above) and bar charts represent number of top intronic mutated genes (below).

d, Ti/Tv ratio of the variant analysis of primary PDAC and their matching models.

e, Shared intronic variants and their classifications among all primary PDAC: TM16, TM27 and TM 56. Number of totals discovered and quality filtered intronic mutations from each sample (left), distribution of variant types among shared (78K) intronic mutations (middle), pie charts showing the distribution and location of the intronic variants in intronic, intergenic and coding regions, and a pie chart showing the intronic variants detected on coding and non-coding transcripts(right).

Fig. 3: Selection and classification of genes exhibiting highest numbers of intronic variants.

a, Boxplot showing the distribution of intron length and number of introns of top 20 genes with intronic mutations.

b, Mutation percentages of all discovered mutations based on the variant locations (SS, Prox, Deep) and enriched pathway network representation with identified mutations.

c, Read coverage of snRNA-seq and bulk RNA-seq data visualizing the intronic variant's location for *WWOX* and *SMYD3.*

Fig. 4: Tumor subtype and marker identification at the single-nucleus level.

a, UMAP clustering of single cells from the three patients colored by tumor model (left) and by inferred tumor subtype (right).

b, UMAP clustering showing the gene expression of the top transcriptional markers comparing classical vs basal-like cells.

c, UMAP clustering showing the mutated primary tumor cells for *CNTNAP2*, *CSMD1* and *RBFOX1*

Fig. 5: Identification of markers for BL and CL tumor cells.

a, Heatmap showing gene expression of the top 15 basal-like and classical-like markers overexpressed in all PDAC. Venn diagrams illustrate the overlap of marker genes between tumor PDAC patient samples.

b, Dot plot showing expression of the markers in each of the pancreas cell subtypes. Circle size is proportional to the percentage of cells in each cell type expressing the marker and circle color represents the average marker gene expression in the cell type.

c, UMAP plots by patient showing clusters annotated to specific cell subtypes. The violin plots below represent five classical pancreas markers across cell subtypes.

d, Table showing the number of intronic mutations of the top CL and BL marker genes.

Fig. 6: Combinatorial splicing events related to intronic variants in PDAC.

a, Partial exon skipping with pseudo-exon activation in *CSMD1* (chr8: 3,187,000-3,202,000). 1: partial exon skipping in *CSMD1* (chr8: 3,189,980-3,190,067); 2: pseudo-exon activation in *CSMD1* (chr8: 3,197,495-3,197,578); 3: pseudo-exon activation in the *CSMD1* (chr8: 3,200,391-3,200,466).

b, Four combinatorial IR events in *HNRNPH1* (Chr5: 179,613,890-179,623,583).

Fig. 7: Analysis of IR related to intronic variants in PDAC.

a, Scatter plot of IR ratios from CRU5002 bulk RNA-seq data with their corresponding normalized read depths. Filtered and shared IR events between public PDAC and CRU-5002 data (blue). IR events of high intronic mutated genes (Total-SS-Proximal; green). List of genes exhibiting IR events related to high intronic mutations (right).

b, Kaplan-Meier survival analysis generated via the cBioPortal shows overall survival time of PDAC patients with or without mutations in the top 20 IR genes shared by all three tumors.

c, Region of high percentage IR event detected by IRFinder for the *INTS3* gene shared in KFO and public PDAC (Mean IRraio: 0.45).

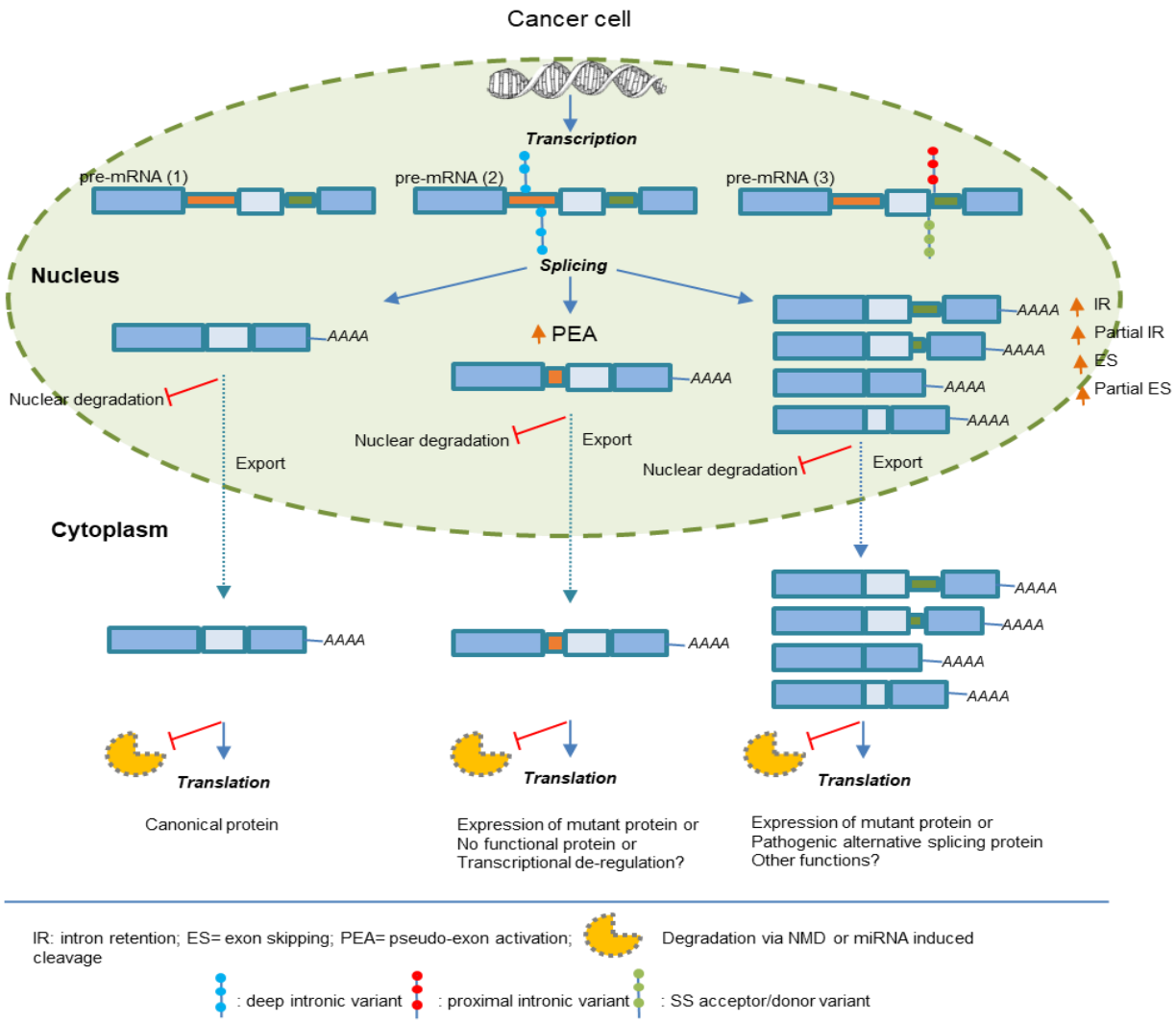
d, Second IR event in the *INTS3* gene (Mean IRraio: 0.25).

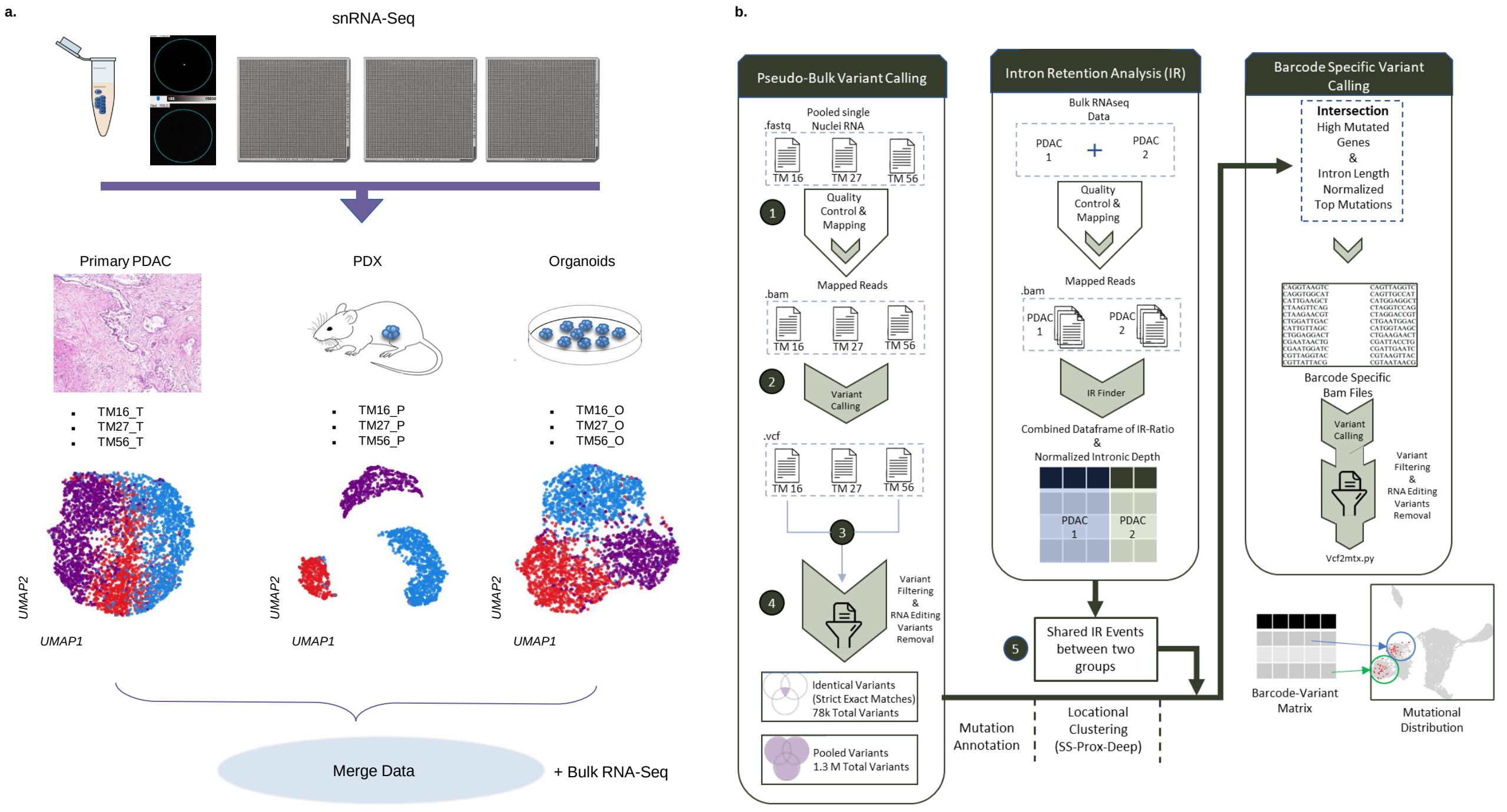
e, SR binding motifs identified in *INTS3*. Two CAGCAGG binding sites for *SRSF1* and one CTCCCGG motif for SRSF2.

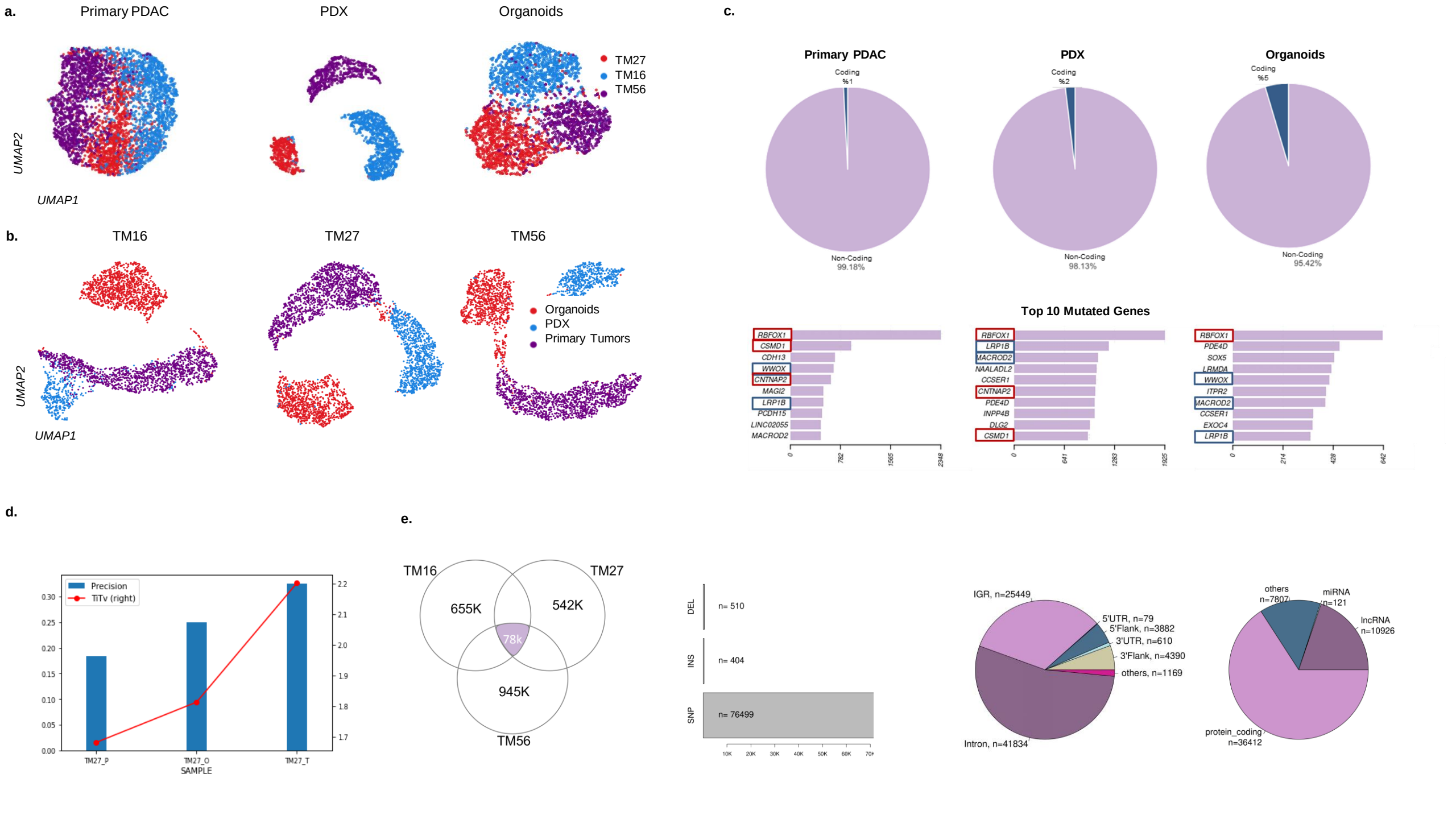
Table 1: Classification of high intronic-mutated genes based on variant location.

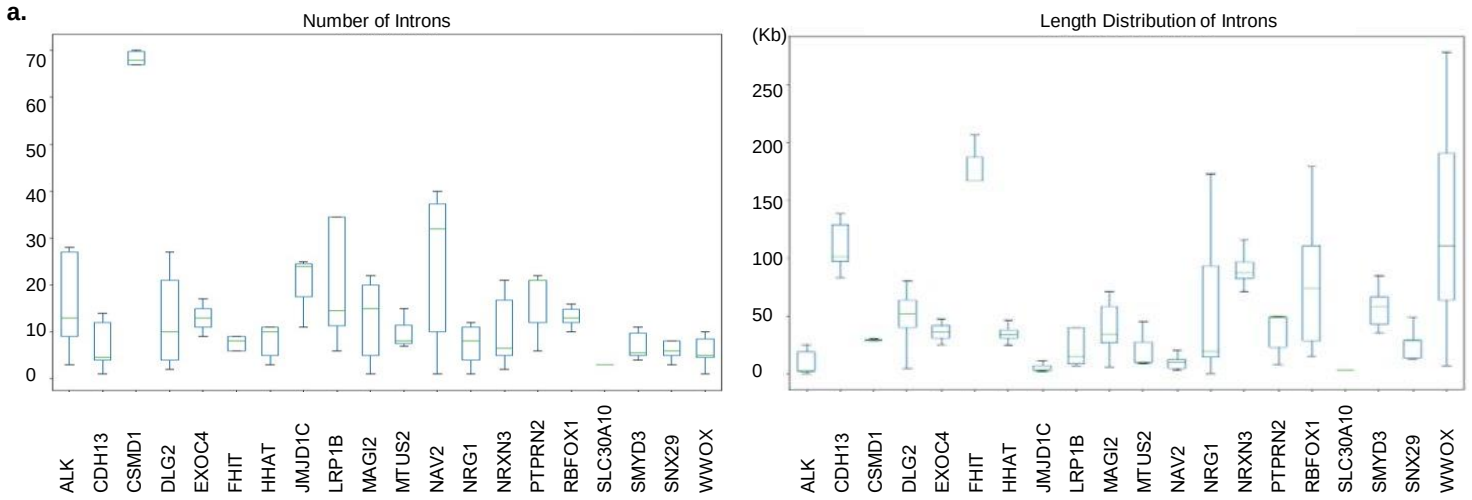
Classification of highly intronic mutated genes based on the variant location: donor and acceptor splice sites (SS, 1–2 bp away from the nearest exon-intron junction); proximal site (3–20 bp away from the nearest exon-intron junction), and deep intronic (>20 bp away from the nearest exon-intron junction). Total number of intronic mutations based on genes described in Supplementary Table S2.

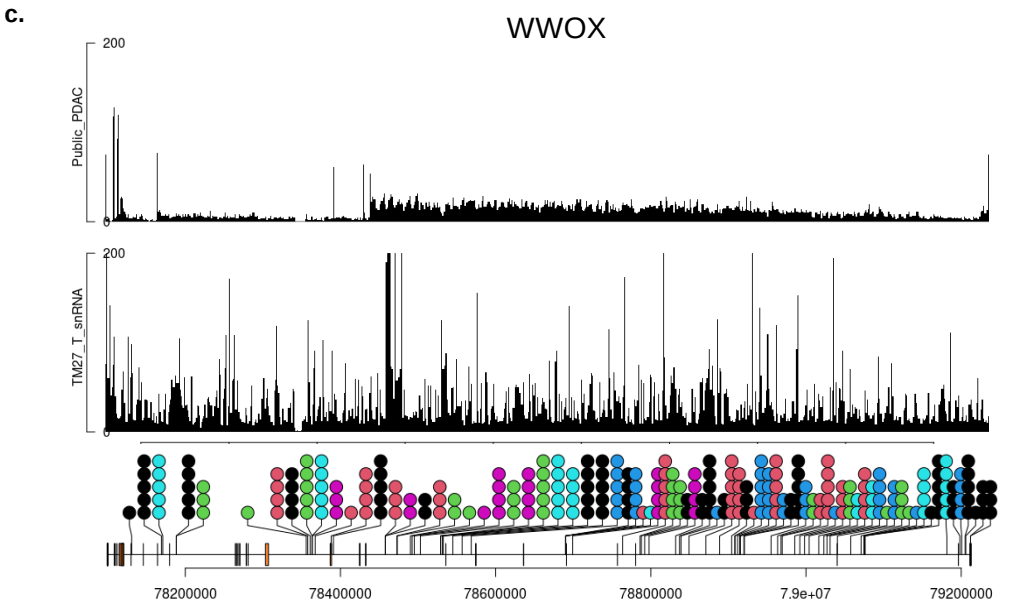
Abstract Graph

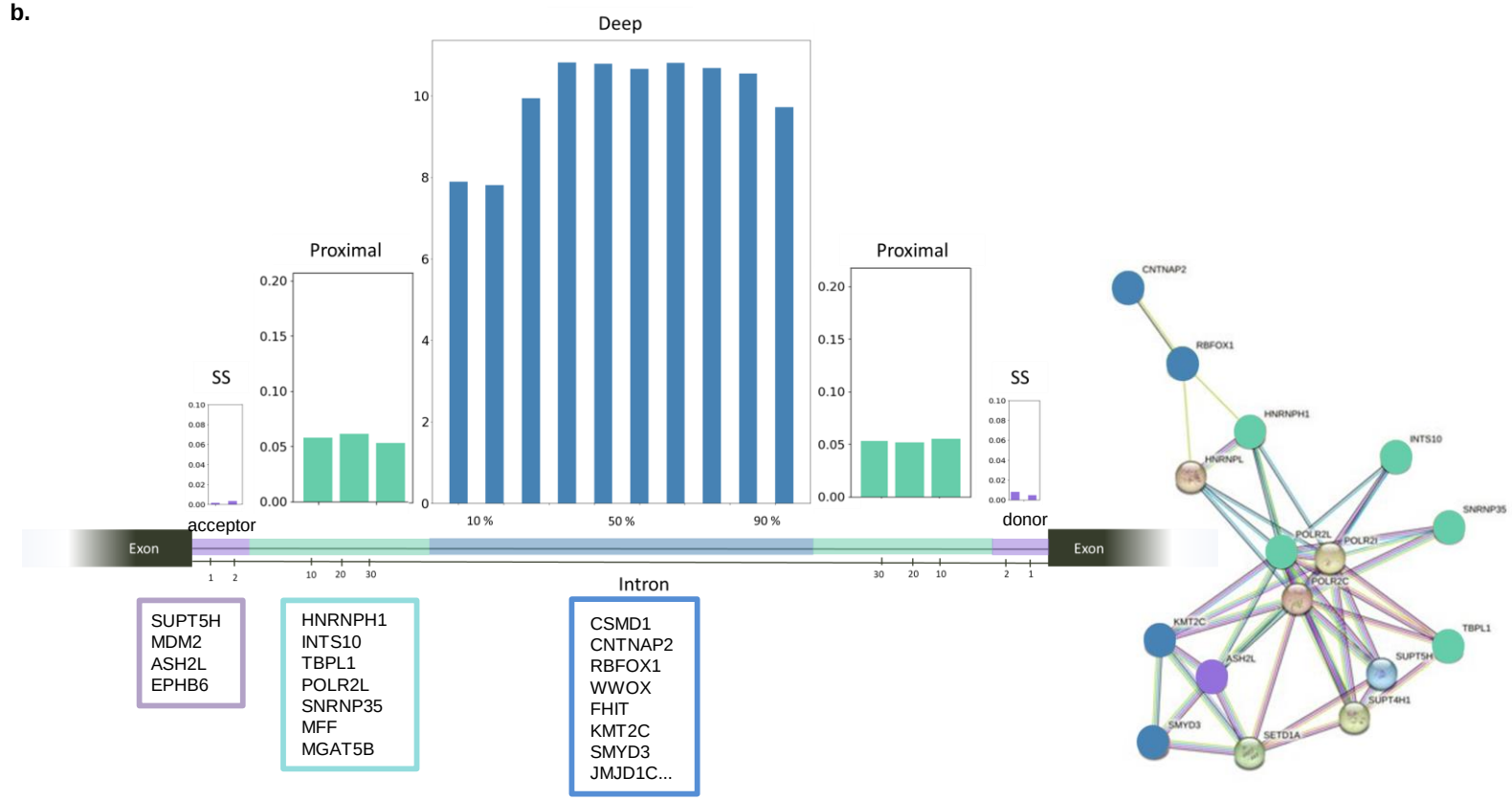


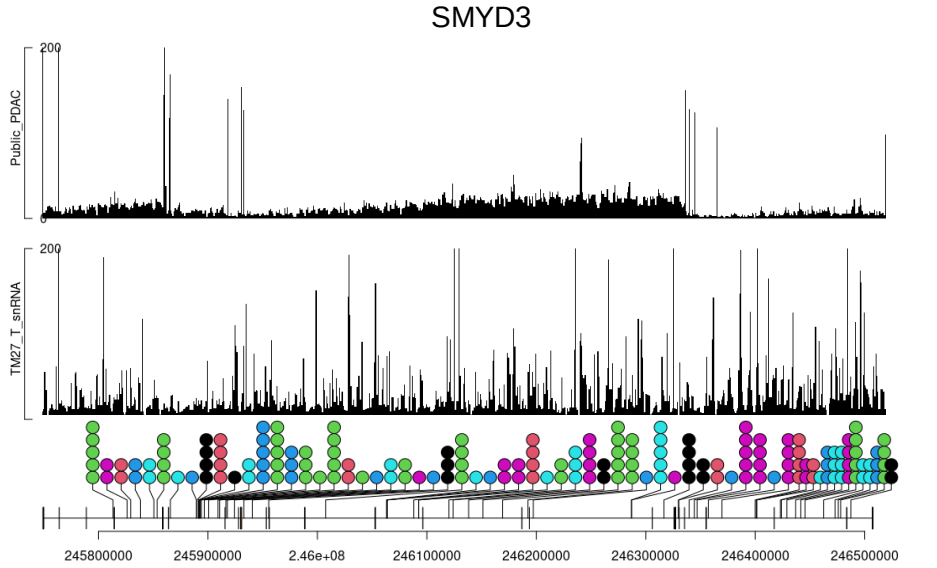

a.
snRNA-Seq
Primary PDAC
PDX
Organoids
TM16_T
TM27_T
TM56_T
TM16_P
TM27_P
TM56_P
TM16_O
TM27_O
TM56_O
UMAP2
UMAP1
Merge Data
+ Bulk RNA-Seq
b.
Pseudo-Bulk Variant Calling
Pooled single Nuclei RNA
.fastq
TM 16
TM 27
TM 56
Quality Control & Mapping
Mapped Reads
.bam
Variant Calling
.vcf
Variant Filtering & RNA Editing Variants Removal
Identical Variants (Strict Exact Matches) 78k Total Variants
Pooled Variants 1.3 M Total Variants
Intron Retention Analysis (IR)
Bulk RNAseq Data
PDAC 1
PDAC 2
IR Finder
Combined Dataframe of IR-Ratio & Normalized Intronic Depth
Shared IR Events between two groups
Mutation Annotation
Locational Clustering (SS-Prox-Deep)
Barcode Specific Variant Calling
Intersection
High Mutated Genes & Intron Length Normalized Top Mutations
Barcode Specific Bam Files
Vcf2mtx.py
Barcode-Variant Matrix
Mutational Distribution


a.
Primary PDAC
PDX
Organoids
TM27
TM16
TM56
UMAP2
UMAP1
b.
TM16
TM27
TM56
Organoids
PDX
Primary Tumors
UMAP2
UMAP1
c.
Primary PDAC
Coding %1
Non-Coding 99.18%
PDX
Coding %2
Non-Coding 98.13%
Organoids
Coding %5
Non-Coding 95.42%
Top 10 Mutated Genes
RBFOX1
CSMD1
CDH13
WWOX
CNTNAP2
MAGI2
LRP1B
PCDH15
LINC02055
MACROD2
0
782
1565
2348
RBFOX1
LRP1B
MACROD2
NAALADL2
CCSER1
CNTNAP2
PDE4D
INPP4B
DLG2
CSMD1
0
641
1283
1925
RBFOX1
PDE4D
SOX5
LRMDA
WWOX
ITPR2
MACROD2
CCSER1
EXOC4
LRP1B
0
214
428
642
d.
Precision
TiTv (right)
0.30
0.25
0.20
0.15
0.10
0.05
0.00
2.2
2.1
2.0
1.9
1.8
1.7
TM27_P
TM27_O
TM27_T
SAMPLE
e.
TM16
TM27
TM56
655K
542K
78k
945K
DEL
n= 510
INS
n= 404
SNP
n= 76499
10K
20K
30K
40K
50K
60K
70K
IGR, n=25449
5'UTR, n=79
5'Flank, n=3882
3'UTR, n=610
3'Flank, n=4390
others, n=1169
Intron, n=41834
others n=7807
miRNA n=121
lncRNA n=10926
protein_coding n=36412


a.
Number of Introns
Length Distribution of Introns
(Kb)
ALK
CDH13
CSMD1
DLG2
EXOC4
FHIT
HHAT
JMJD1C
LRP1B
MAGI2
MTUS2
NAV2
NRG1
NRXN3
PTPRN2
RBFOX1
SLC30A10
SMYD3
SNX29
WWOX


c.
WWOX
Public_PDAC
TM27_T_snRNA
78200000
78400000
78600000
78800000
7.9e+07
79200000

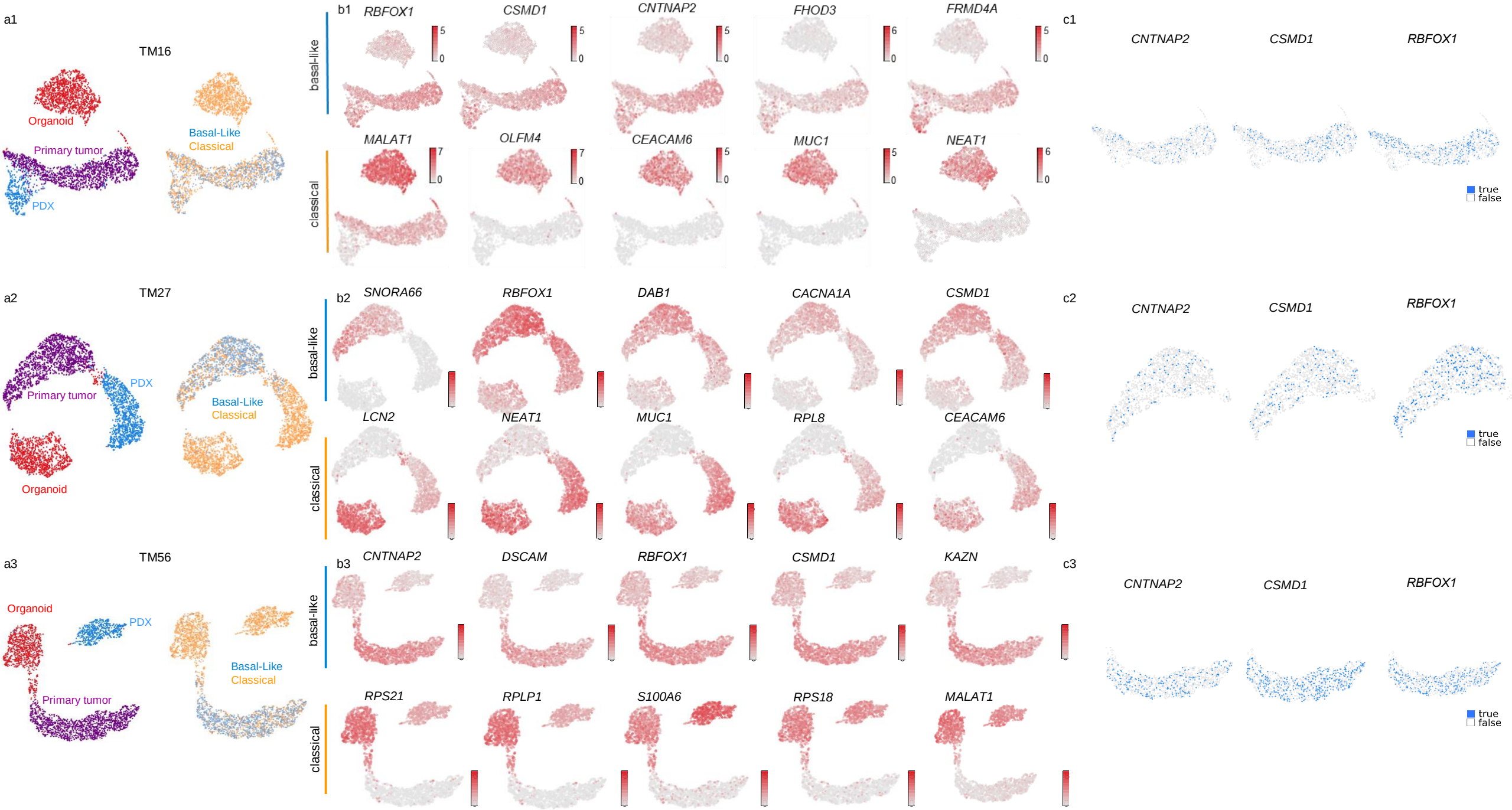

b.
Deep
Proximal
Proximal
SS
SS
acceptor
donor
Exon
Exon
Intron
10 %
50 %
90 %
SUPT5H
MDM2
ASH2L
EPHB6
HNRNPH1
INTS10
TBPL1
POLR2L
SNRNP35
MFF
MGAT5B
CSMD1
CNTNAP2
RBFOX1
WWOX
FHIT
KMT2C
SMYD3
JMJD1C...

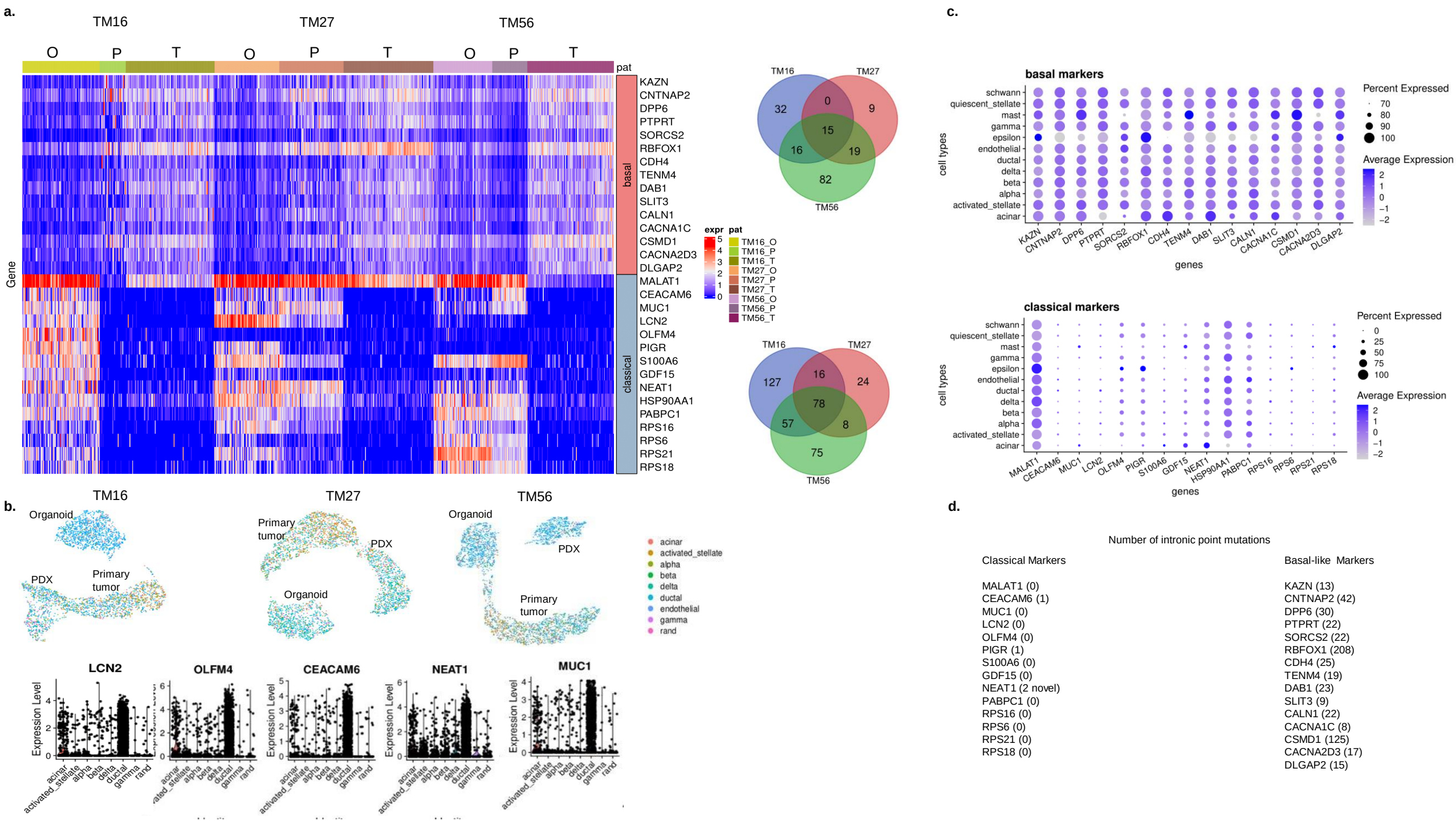

SMYD3
Public_PDAC
TM27_T_snRNA
245800000
245900000
2.46e+08
246100000
246200000
246300000
246400000
246500000

a. Subtyping classical and basal-like cells

b. Transcriptional markers: classical and basal-like cells

c. Intronic variants distribution of basal-like markers in primary PDAC

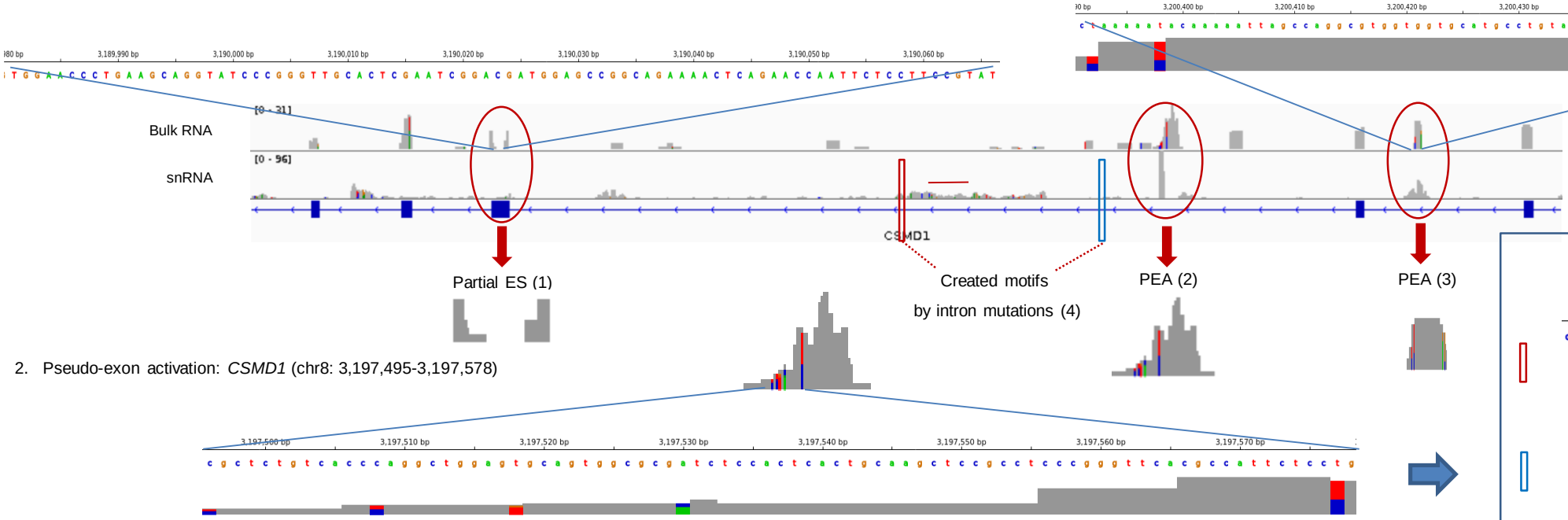



a.
TM16
TM27
TM56
O
P
T
pat
Gene
basal
classical
KAZN
CNTNAP2
DPP6
PTPRT
SORCS2
RBFOX1
CDH4
TENM4
DAB1
SLIT3
CALN1
CACNA1C
CSMD1
CACNA2D3
DLGAP2
MALAT1
CEACAM6
MUC1
LCN2
OLFM4
PIGR
S100A6
GDF15
NEAT1
HSP90AA1
PABPC1
RPS16
RPS6
RPS21
RPS18
expr
pat
TM16_O
TM16_P
TM16_T
TM27_O
TM27_P
TM27_T
TM56_O
TM56_P
TM56_T
32
0
9
15
16
19
82
127
16
24
78
57
8
75
b.
Organoid
PDX
Primary tumor
acinar
activated_stellate
alpha
beta
delta
ductal
endothelial
gamma
rand
LCN2
OLFM4
CEACAM6
NEAT1
MUC1
Expression Level
c.
basal markers
classical markers
cell types
genes
Percent Expressed
Average Expression
d.
Number of intronic point mutations
Classical Markers
MALAT1 (0)
CEACAM6 (1)
MUC1 (0)
LCN2 (0)
OLFM4 (0)
PIGR (1)
S100A6 (0)
GDF15 (0)
NEAT1 (2 novel)
PABPC1 (0)
RPS16 (0)
RPS6 (0)
RPS21 (0)
RPS18 (0)
Basal-like Markers
KAZN (13)
CNTNAP2 (42)
DPP6 (30)
PTPRT (22)
SORCS2 (22)
RBFOX1 (208)
CDH4 (25)
TENM4 (19)
DAB1 (23)
SLIT3 (9)
CALN1 (22)
CACNA1C (8)
CSMD1 (125)
CACNA2D3 (17)
DLGAP2 (15)

a. Combinatorial splicing events: Partial exon skipping with pseudo-exon activation in the CSMD1 transcript (chr8: 3,187,000-3,202,000)

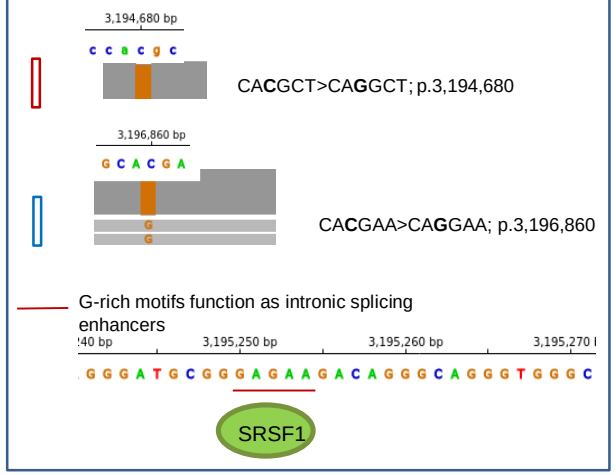
1. Partial exon skipping: *CSMD1* (chr8: 3,189,980-3,190,067)

3. Pseudo-exon activation: *CSMD1* (chr8: 3,200,391-3,200,466)

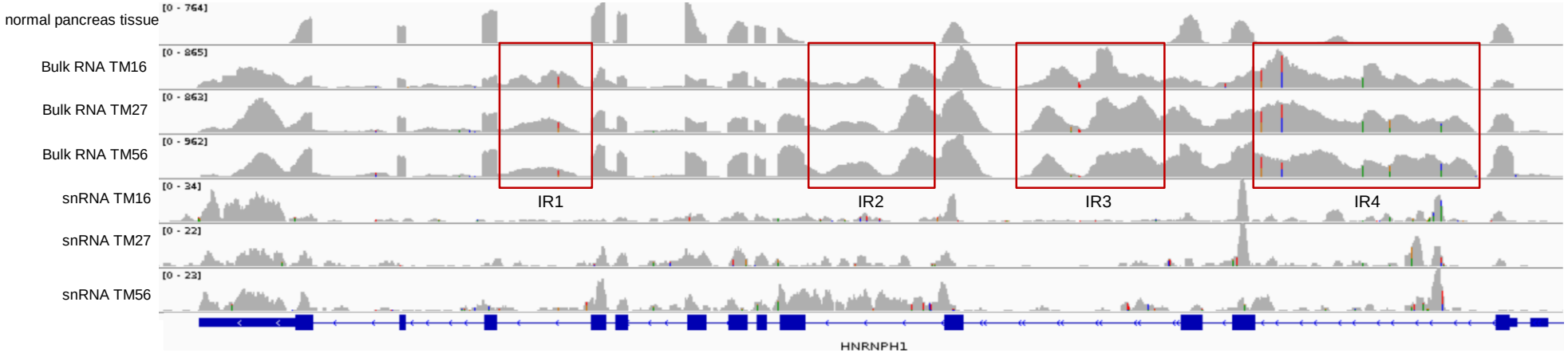


2. Pseudo-exon activation: *CSMD1* (chr8: 3,197,495-3,197,578)

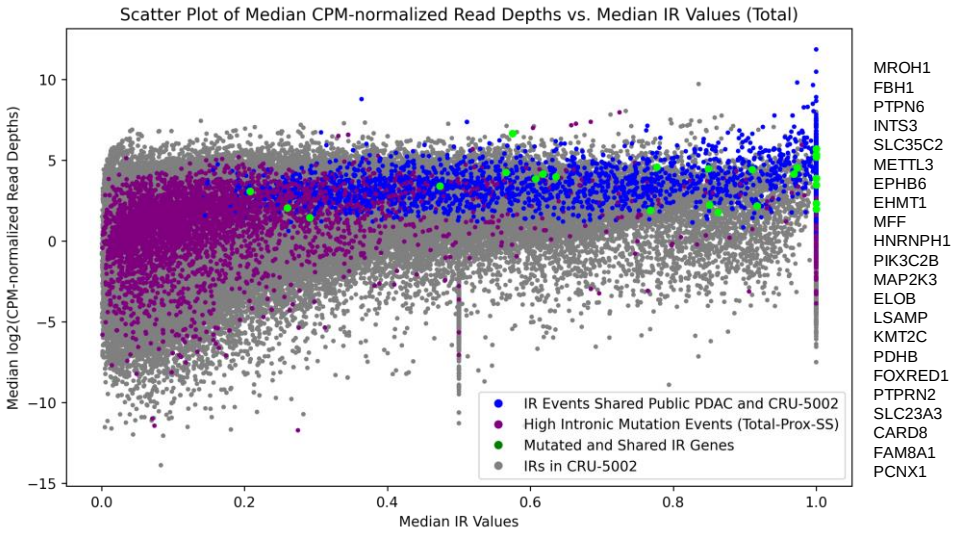
4. Created SR motifs by intronic mutations: *SRSF2* motif 3 [35]



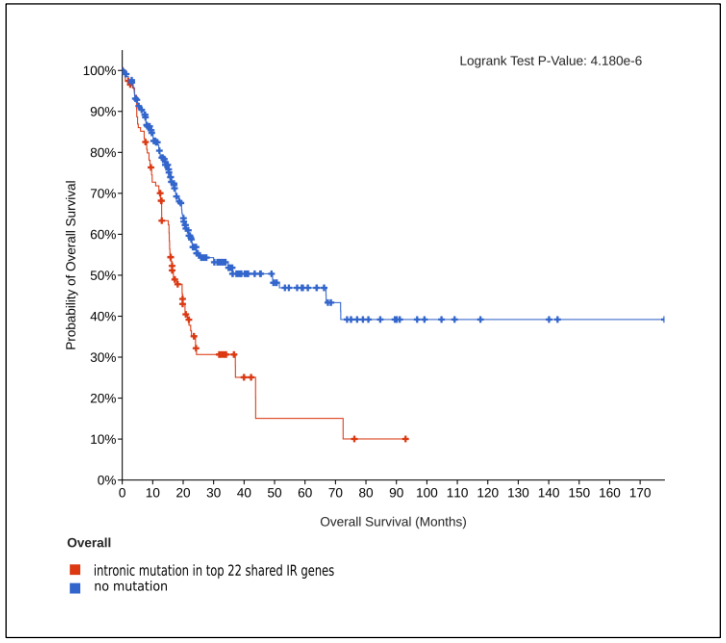
b. Combinatorial splicing events: four total IR in the *HNRNPH1* transcript (Chr5: 179,613,890-179,623,583)



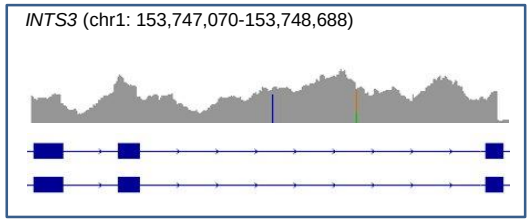
a. Genes showing common IR events and intronic mutations



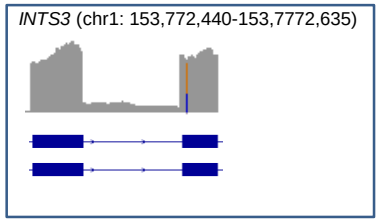
b.



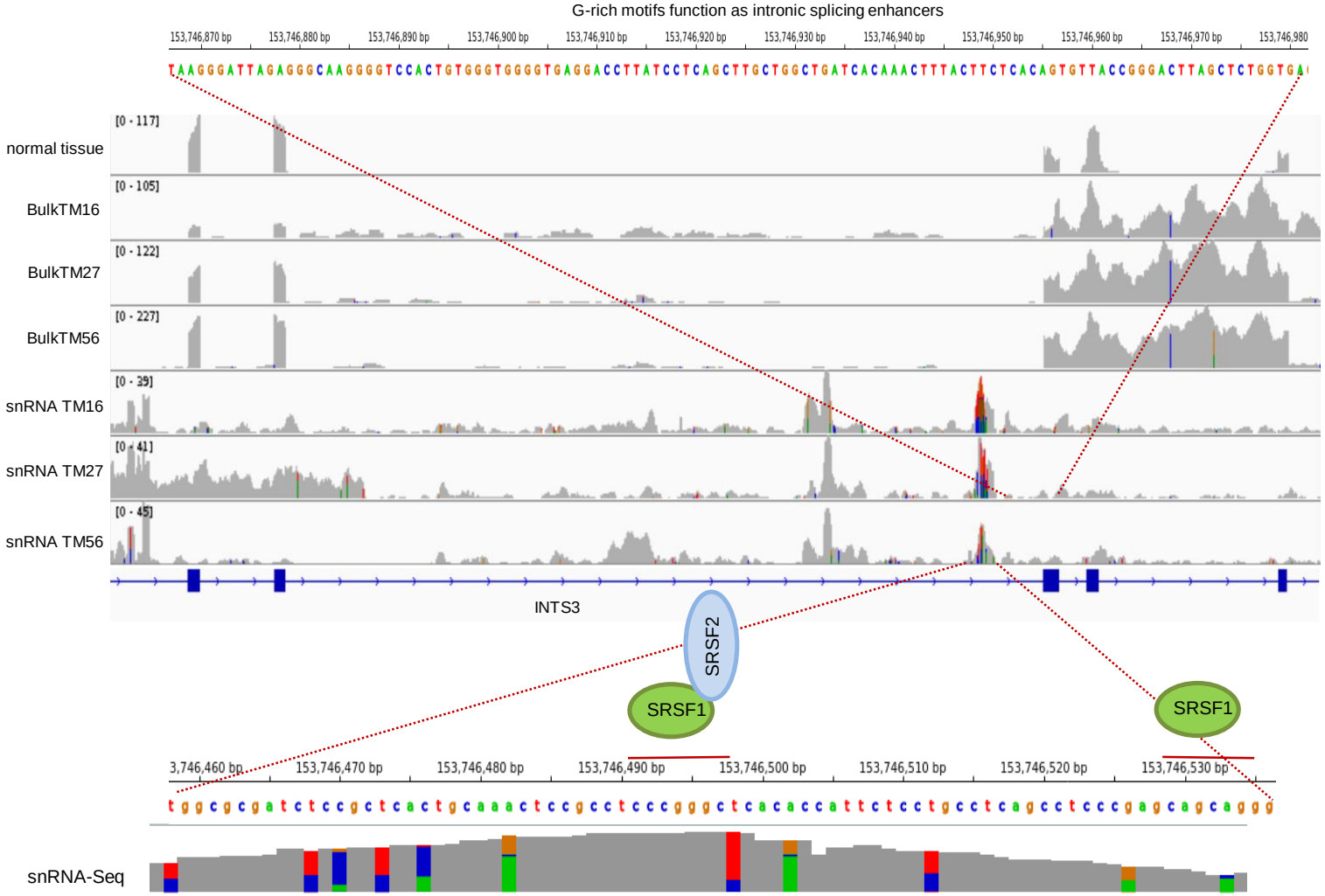
c. Total IR1: INTS3



d. Total IR2: INTS3

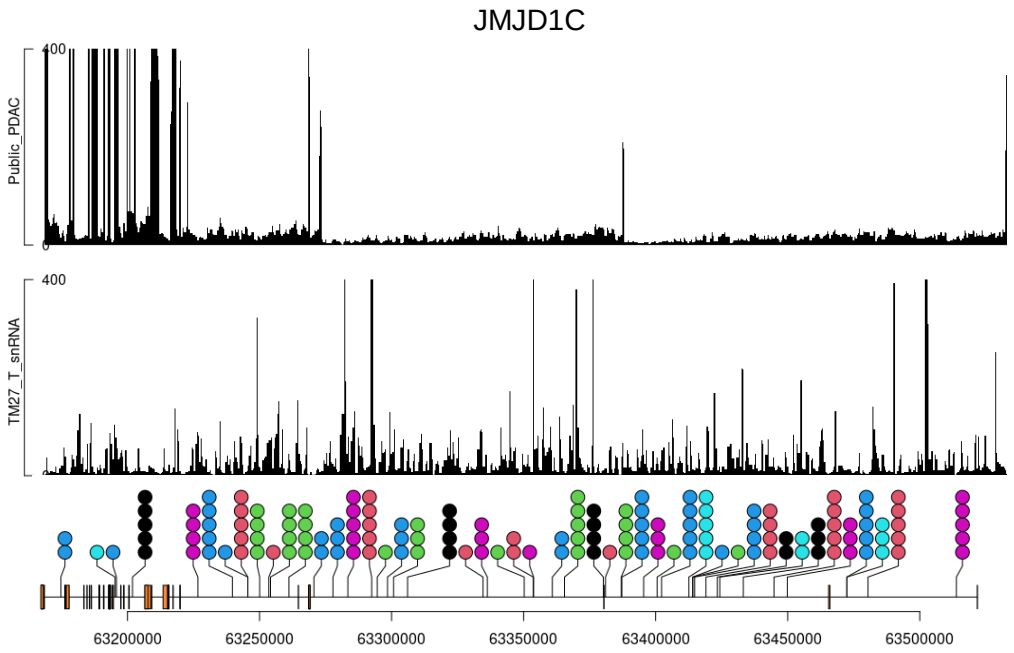


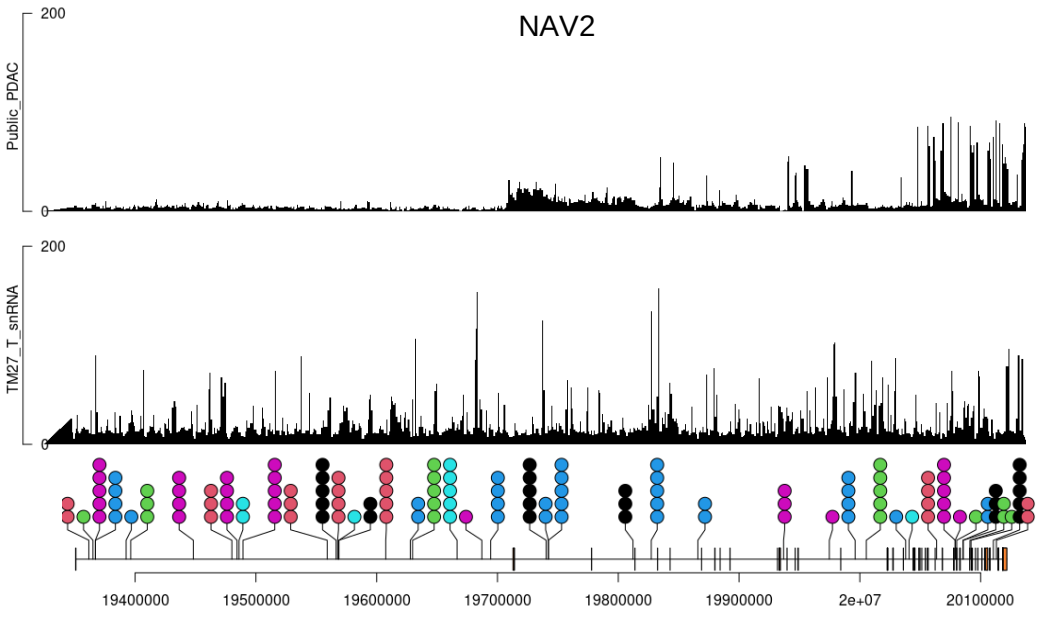
e. SR, GGG- and created CCC- rich motifs identified for IR1: *INTS3* gene, chr1: 153,747,070-153,748,688



Supplementary Figure S1

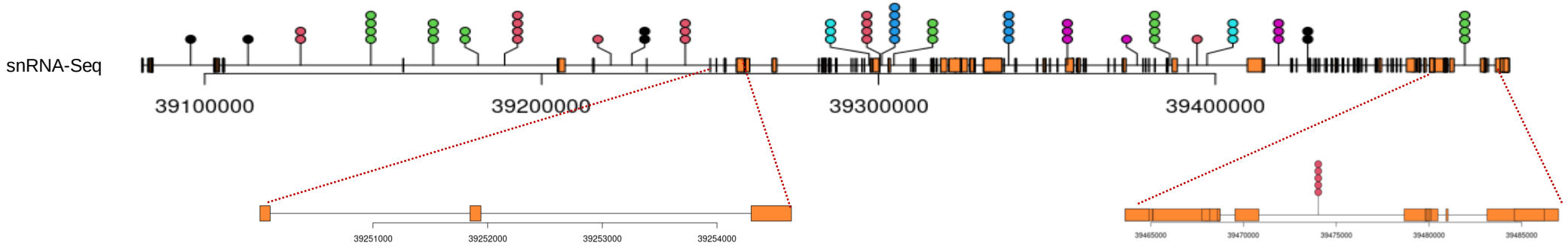
Intronic variant distribution and location for NAV2 and JMJCD1

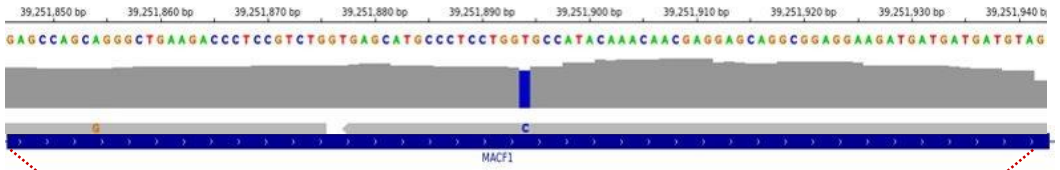




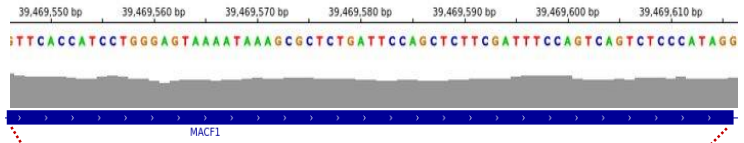
Supplementary Figure S2

Total ES events in normal and tumor pancreas tissue generating a new isoform of MACF1 (chr1:39,081,316-39,487,177)

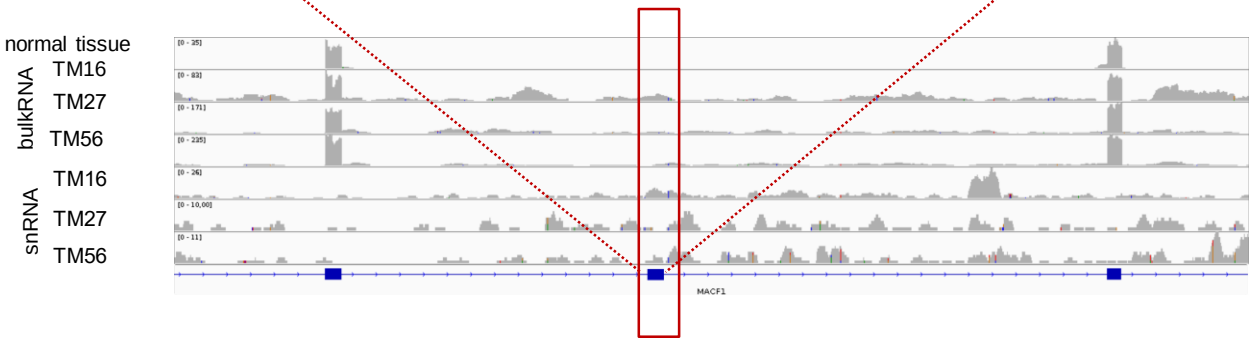




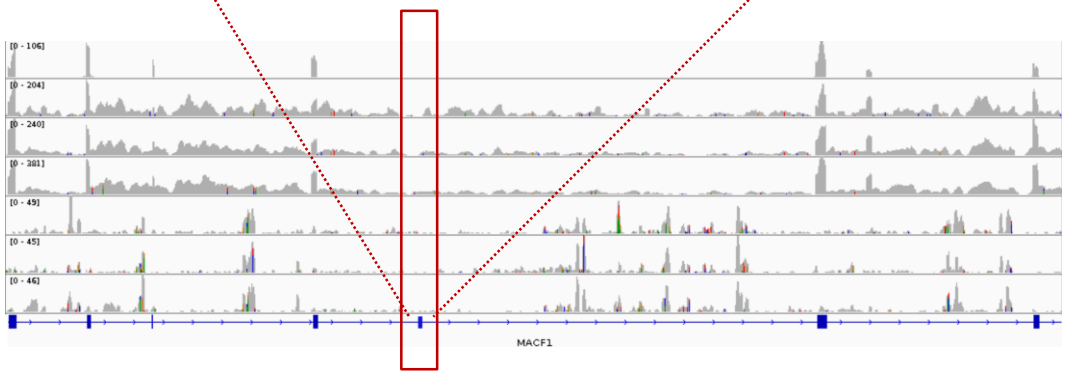
a. Total exon skipping: MACF1 (chr1: 39,251,846-39,251,941)



b. Total exon skipping: MACF1 (chr1: 39,469,547-39,469,615)



Total ES (a)

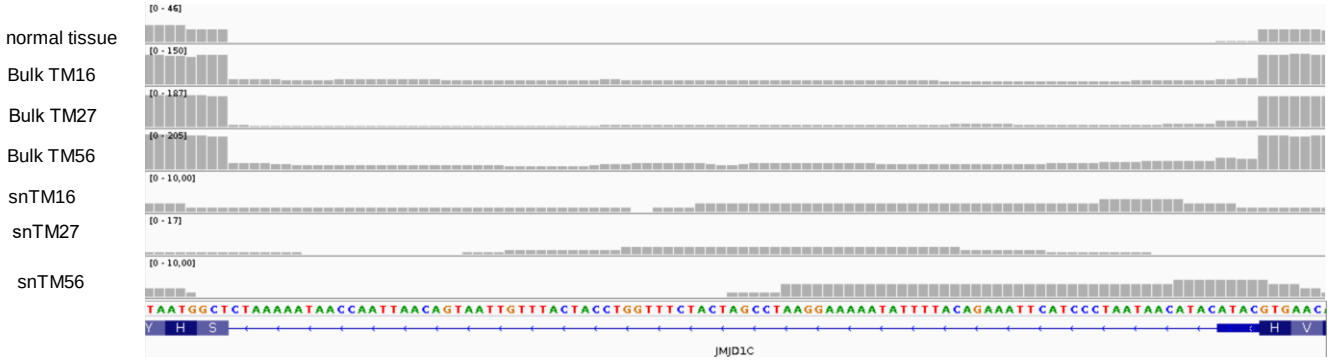
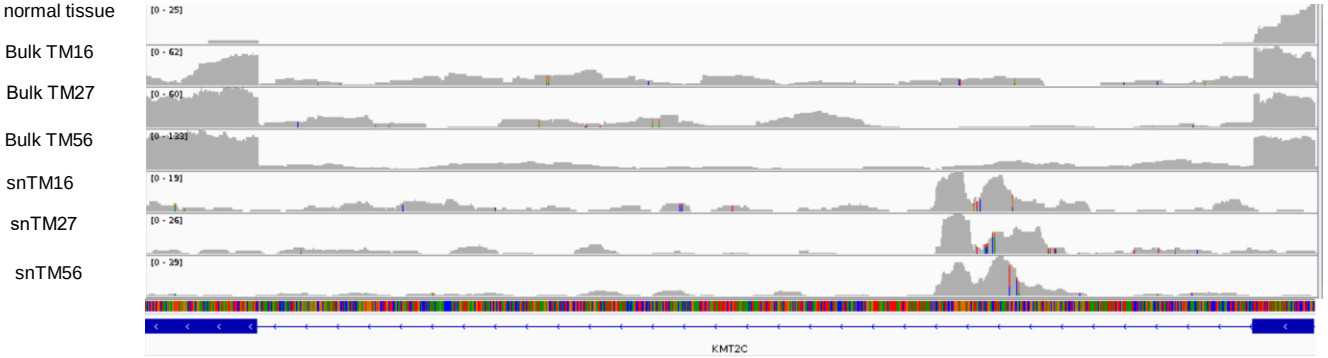


Total ES (b)

Supplementary Figure S3

IR events in primary PDAC: JMJD1C and KMT2C

JMJD1C (Chr10: 63,215,448-63,215,559)



KMT2C (Chr7: 152,148,954-152,151,024)