

# Supplementary data

## Note A: Alignment parameters

### uLTRA

We ran uLTRA (v 0.0.1, commit 3474f3c9) for all five datasets.

#### Indexing

We ran indexing, which consists of two steps `prep\_splicing` and `prep\_seqs` as follows.

```
uLTRA prep_splicing --disable_infer GTF_annotation outfolder
```

```
uLTRA prep_seqs ref_fasta outfolder --min_mem X
```

Where X=17 for DROS, SIRV, and the simulated read datasets, and X=20 for ALZ and ENS data. Also, the prep\_splicing method sets the following default parameters: --flank\_size 1000, --mask\_threshold 200, --small\_exon\_threshold 200.

#### Alignment

Within the algorithm, uLTRA sets the --ont parameter for the DROS, SIM\_ANN, and SIM\_NIC datasets to encompass an error rate of 5-12%, and --isoseq parameter for the SIM\_ENS and ALZ dataset for lower error rates. The --ont parameter sets minimum MEM size of 17 and MAM alignment minimum accuracy of 0.6, while --isoseq sets minimum MEM size of 20 and MAM alignment minimum accuracy of 0.8. uLTRA uses slaMEM (Fernandes and Freitas 2014) to find MEMs, edlib (Šošic and Šikic 2017) to find MAMs, and parasail (Daily 2016) to perform final alignment to exons. We use alignment with parasail in semi-global mode with alignment penalties chosen as match:2, mismatch:-2, gap open: -3, gap extend: -1.

uLTRA also sets the following parameters related to the number of alignments to report --dropoff 0.95, --max\_loc 5.

## Minimap2

We ran minimap2 (H. Li 2018) (v2.17-r974-dirty, GitHub commit c9874e2). We used -k 14 for PacBio Alzheimer dataset and -k 13 for the other datasets to index the reference genomes. For alignment with minimap2, we specified parameters --eqx -t 62 -ax splice -k14 -G 500k for the PacBio Alzheimer dataset, and --eqx -t 62 -ax splice -k13 -w 5 -G 500k for the rest of the datasets. The -k and -w parameters were set to 13 and 5, respectively, to improve accuracy over default parameters. We observed a 0.25%, 0.5%, and 0.72% increase in the number of correct alignments over default parameters on SIM\_ENS, SIM\_ANN, and SIM\_NIC. We also increased the -G parameter from the default value of 200,000 to 500,000 in all datasets except for SIRV. The parameter -G controls maximum intron length. We observed that minimap2 missed valid high quality mapped isoforms in the ALZ dataset with introns larger than 200,000 (e.g., 10 reads that uLTRA found to transcript ENST00000555571.5 with an intron of length 387,219nt). Finally, minimap2 sets a special penalty score for non-canonical junctions. We set the non-default parameters --splice-flank=no --secondary=no -C5 on the SIRV dataset as suggested in the minimap2 documentation and discussed in issue 99 in the minimap2 repository since SIRV isoforms do not honor the canonical AG-GT junction to the same extent as biological data. We observed a drastic improvement with setting --splice-flank=no compared to default splicing mode. With original splicing parameters, minimap2 only mapped 725860 reads as FSM to 45 unique isoforms.

## deSALT

We ran deSALT (v1.5.5, commit f4648d8) on all datasets besides the ENS and SIM\_ANNOT datasets for which this version did not complete (took more than 96h). On these two datasets, we ran an earlier version (v1.5.1). We indexed the genomes with default parameters, as suggested in (Liu et al. 2019). We ran deSALT alignment with parameters -d 10 and -s 2 for all datasets, as (Liu et al. 2019) used these settings for both ONT and PacBio Iso-Seq data. We used the parameter -l 14, which lowers the seeding kmer size from the default of 15 and is supposed to increase accuracy at the cost of runtime. We also increased the --max-intron-len parameter from the default value of 200,000 to 500,000 in all datasets except for SIRV. The parameter --max-intron-len controls maximum intron length, and we observed, similarly to minimap2, that deSALT missed valid isoforms with intron larger than the default value of 200,000. Also, deSALT, similarly to minimap2, set an individual penalty score for non-canonical junctions. We set the parameter --noncan to 4 instead of the default of 9 on the SIRV data, and, similarly to minimap2's results, and we observed a substantially improved accuracy over default parameters. With original splicing parameters, deSALT only mapped 746,604 reads as FSM to 47 unique isoforms. Finally, we specified -t 48 (number of cores) as deSALT allows a maximum of 48 cores.

## Graphmap2

We ran GraphMap2, github commit 02fcef7a. We set `--threads 62` and `-x rnaseq`.

## Note B: Simulated data

### Simulating from annotated transcripts

We considered using already existing read simulators (NanoSim (Yang et al. 2017), DeepSimulator (Y. Li et al. 2018), SimLord (Stöcker, Köster, and Rahmann 2016), and SNaReSim (Faucon et al. 2017) before implementing a transcriptomic read simulator. However, they are genomic read simulators and cannot easily be modified to simulate full-length transcript reads.

In our simulations, we simulate transcripts from 234,207 distinct ENSEMBL cDNA sequences from the human reference. We sample reads uniformly from the transcripts. Each sampled transcript is then subject to sequencing errors. At each position in the transcript, we enter an error state with probability 0.05. If the simulator enters an error state, it chooses a deletion state with probability 0.5, substitution with probability 0.3, and insertion with probability 0.2. In the deletion and insertion states, the length of the insertion and deletion is simulated from a geometric distribution with parameter 0.5. This model gives an average error rate of 8.6%.

### Simulating novel transcripts

We used the GTF annotations for this simulation. We select all genes with four or more exons that have non-overlapping genomic coordinates. For each transcript we simulate from such a gene, we include the first and last exon with probability 1, but any internal exon with probability 0.5. This probabilistic inclusion of exons gives isoforms. We check if the transcript matches an already annotated isoform. If not, we keep it in the simulation. To keep the number of transcripts equal to the annotated number of transcripts for each gene, we simulate as many novel transcripts as the gene had annotated transcripts (if the number of possible combinations permits).

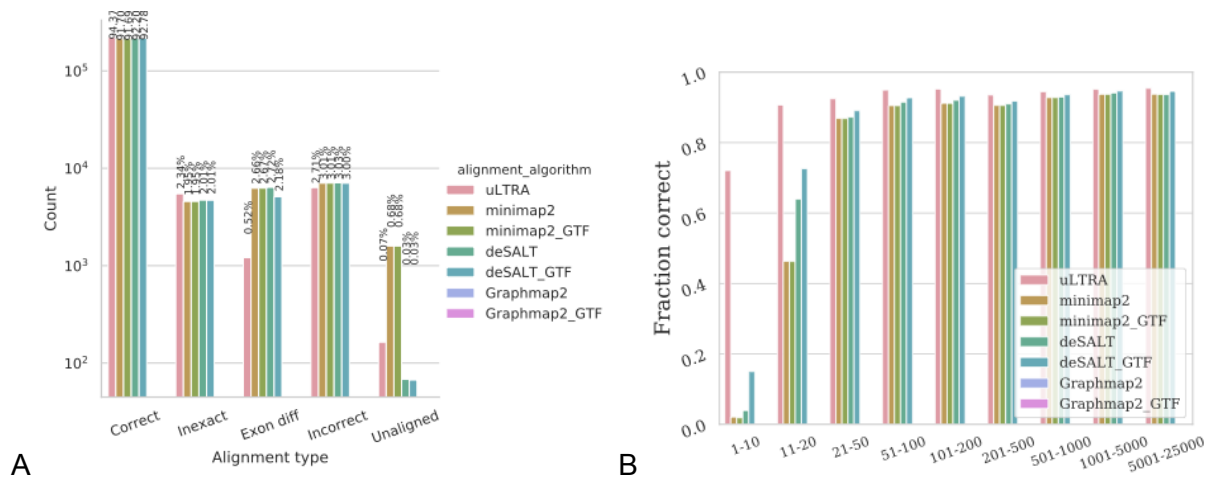
## Evaluation of simulated data

For the simulated data, we have the true genome annotations of each exon in the transcript. A read is therefore classified as *correct* if the read is aligned to all the correct exons, and all of the exon alignments have an offset of fewer than 15 nucleotides to true annotation coordinates. An alignment is *inexact* if the read is aligned to the correct exons, but at least one junction offset is more than 15 nucleotides. An alignment is classified as having an *exon difference* if the read alignment is missing one or more exons, or a segment of at least 15 nucleotides is aligned to a genomic location not included in the set of true exons (i.e., appearing as a false exon). An alignment is classified as an *incorrect location* if it aligns to a genomic location not overlapping with the correct annotation. In our classifications above, we chose 15bp as the threshold because we did not observe deletions or insertions longer than 15 nucleotides in the simulated data. Therefore, an offset larger than 15 nucleotides indicates a misalignment rather than an actual deletion or insertion of this size in the read, causing a junction offset.

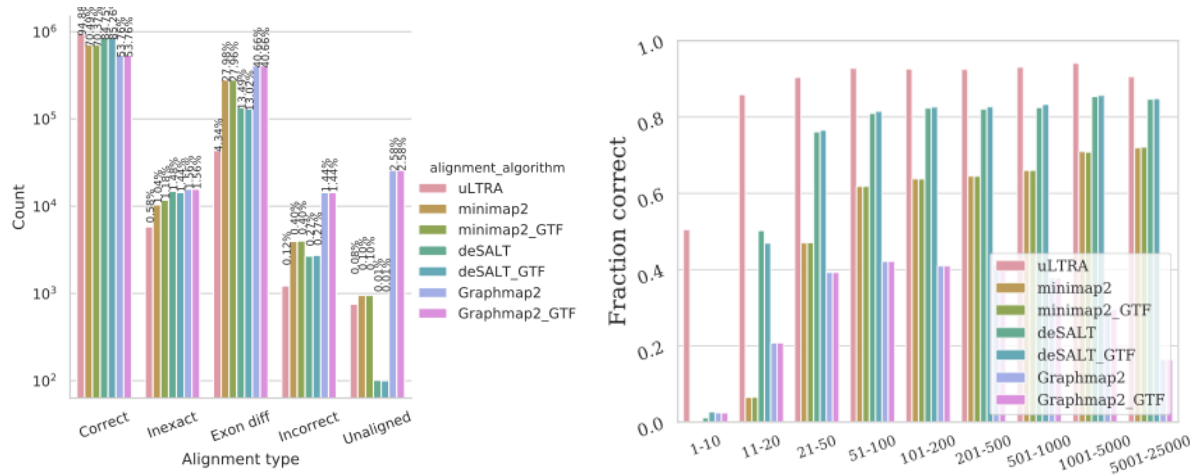
## Note C: SIRV analysis

The SIRV dataset consists of 68 synthetic transcripts from 7 different loci sequenced with ONT R9 technology (see (Sahlin et al., n.d.) for details). The transcripts from each locus differ in their splicing pattern. Eight out of the 68 transcripts contain only one exon, and therefore does not have a splice site. Furthermore, two isoforms SIRV701 and SIRV705 have identical splice sites and differ only in a 2nt offset in both the transcription start and stop site. Therefore we used 59 isoforms with distinct splice sites to investigate alignment performance around splice sites. The ONT SIRV dataset was constructed so that the isoforms should have roughly the same abundance in the sample (the SIRV E0 mix). While sequencing depth bias can cause significant differences in read depth due to the isoforms having different lengths, with the given depth of 1.4 million reads, we expect all the isoforms to occur in the sample.

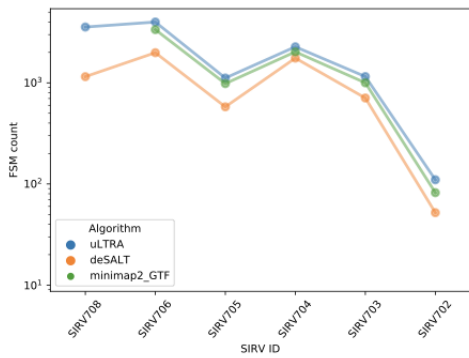
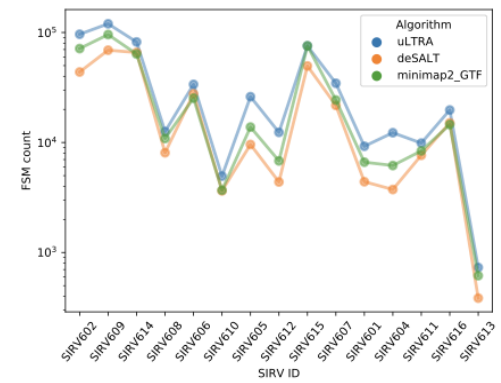
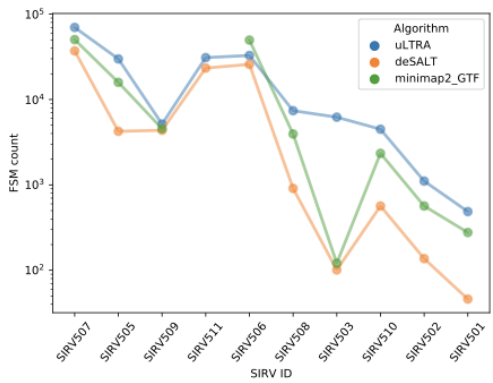
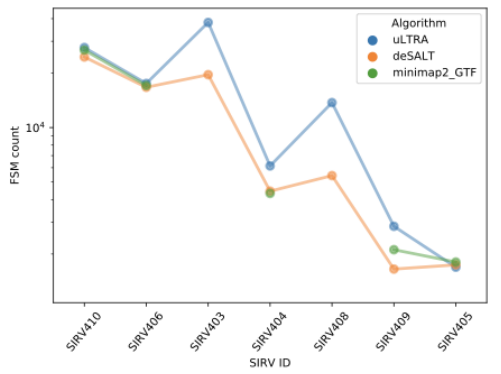
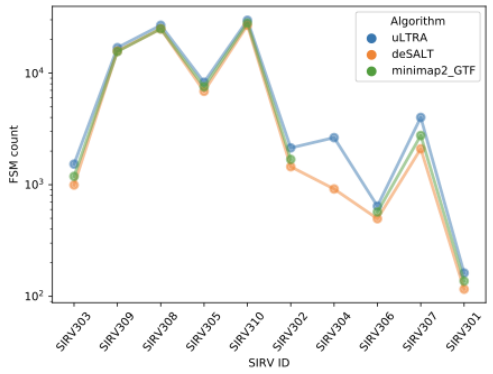
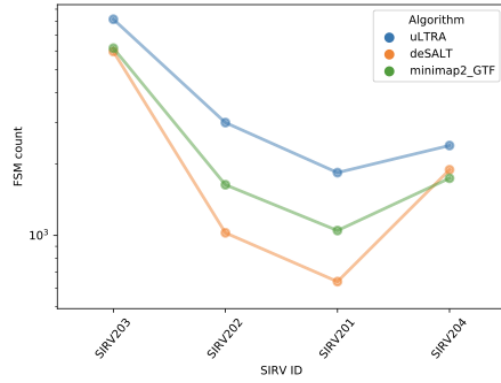
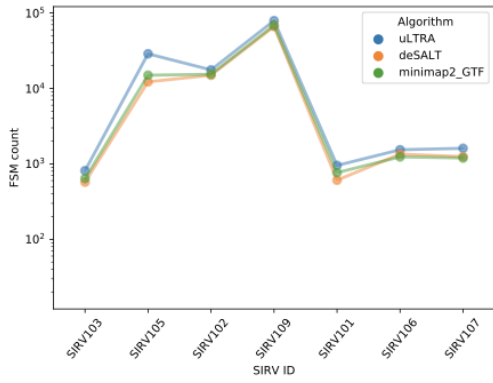
# Supplementary Figures



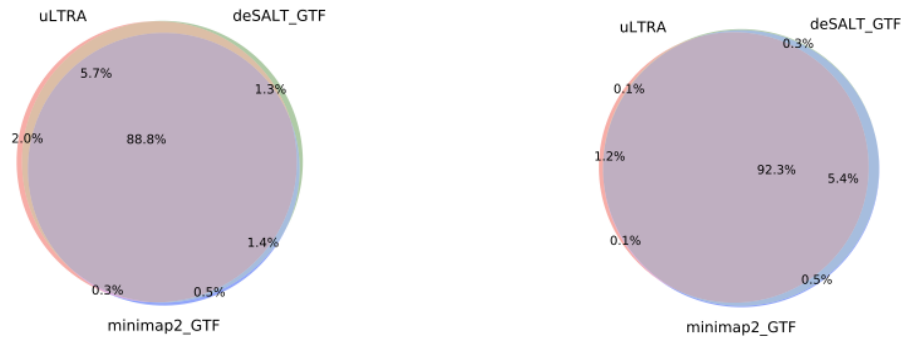
**Figure S1.** Alignment results on simulated data for the ENS dataset. **(A)** Percentage of reads in each respective category. **(B)** The fraction of correctly aligned exons (y-axis) as a function of exon size (x-axis). We could not evaluate GraphMap2 as it encountered an error on the ENS data.



**Figure S2.** Alignment results on simulated data for the SIM\_NIC dataset. **(A)** Percentage of reads in each respective category. **(B)** The fraction of correctly aligned exons (y-axis) as a function of exon size (x-axis). We could not evaluate GraphMap2 as it encountered an error on the ENS data.

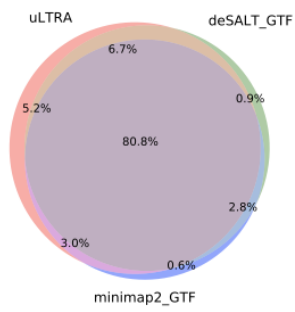


**Figure S3.** The number of reads annotated as FSM to each of the 59 SIRV isoforms with at least one splice site (y-axis log scale). One panel per gene loci.

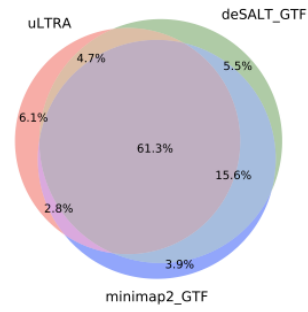


**Figure S4.** Alignment overlap concordance between uLTRA, deSALT\_GTF, and minimap2\_GTF for DROS (A) and ALZ (B).

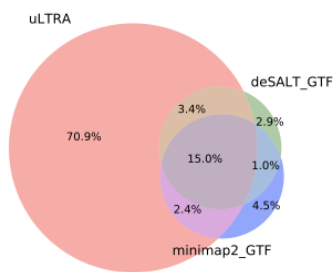




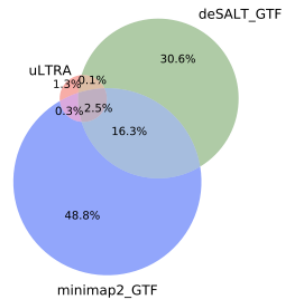
A



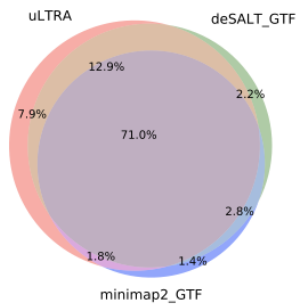
B



C

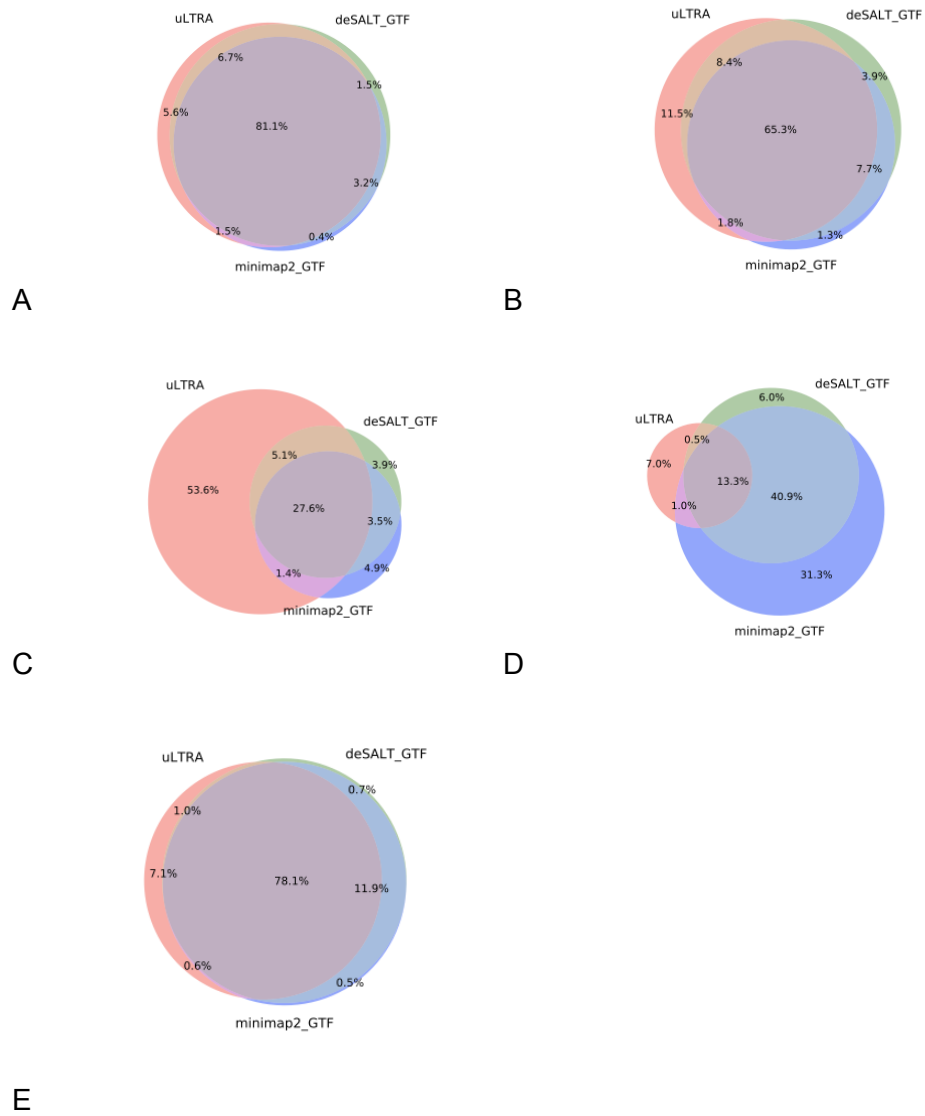


D

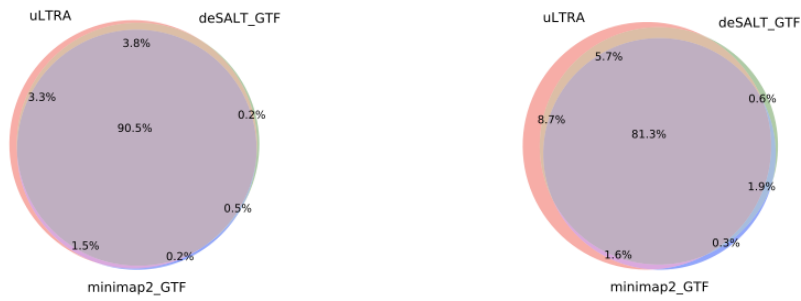


E

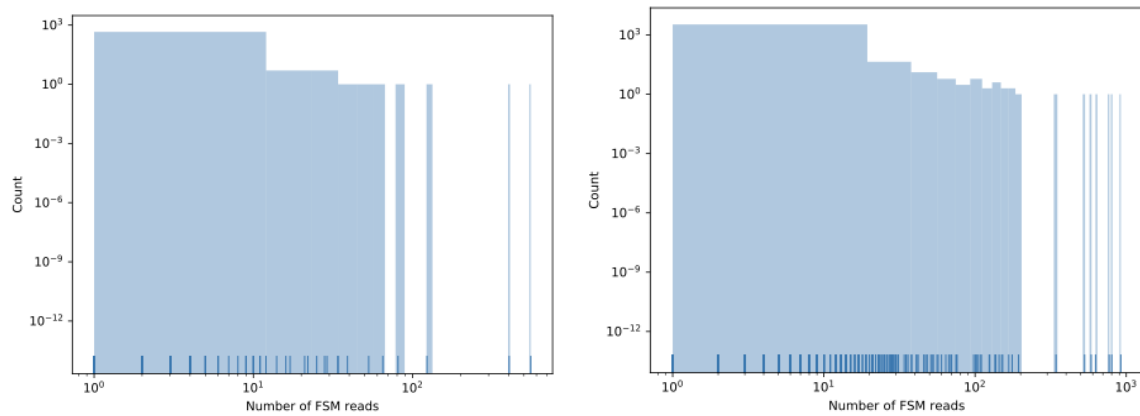
**Figure S5.** Alignment concordance between uLTRA, deSALT\_GTF, and minimap2\_GTF for the different categories FSM (A), ISM (B), NIC (C), NNC (D) and NO\_SPLICE (E) in DROS.



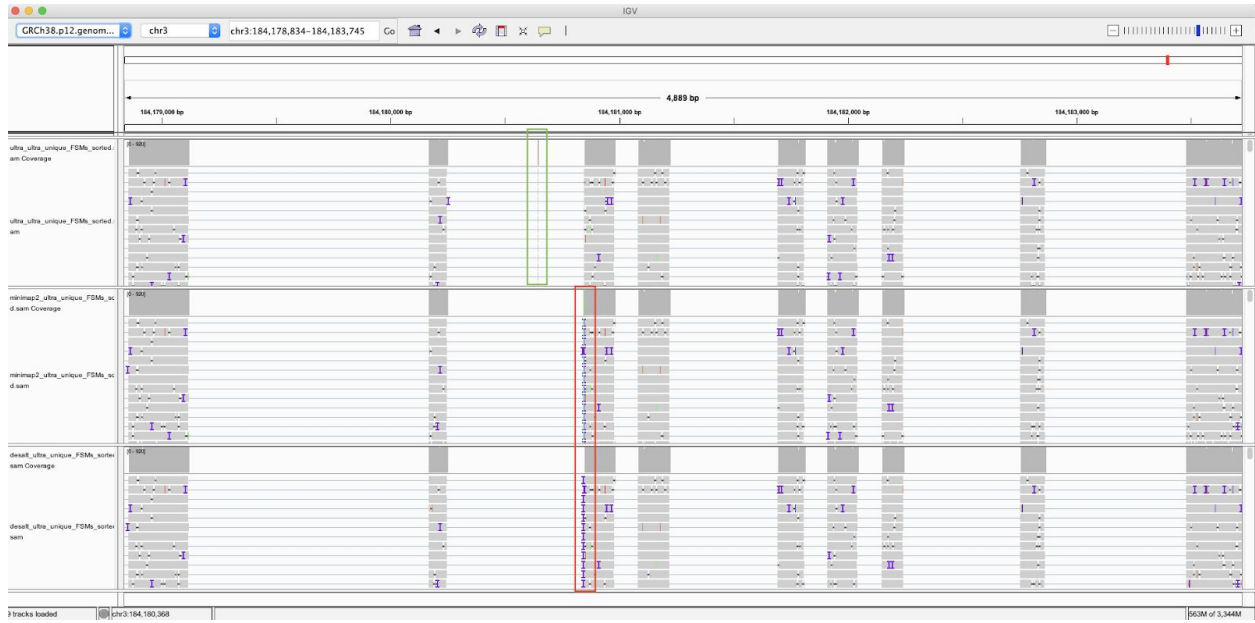
**Figure S6.** Alignment concordance between uLTRA, deSALT\_GTF, and minimap2\_GTF for the different categories FSM (A), ISM (B), NIC (C), NNC (D) and NO\_SPLICE (E) in ALZ.



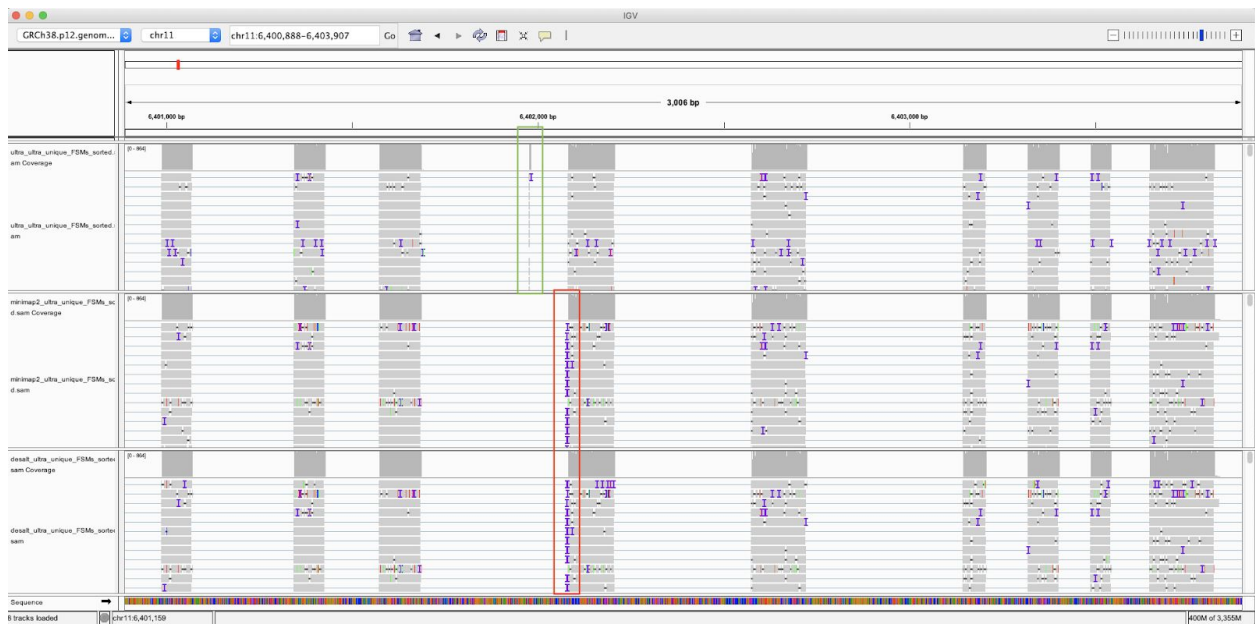
**Figure S7.** Concordance of unique isoforms that had FSM aligned reads for DROS (A) and ALZ (B).



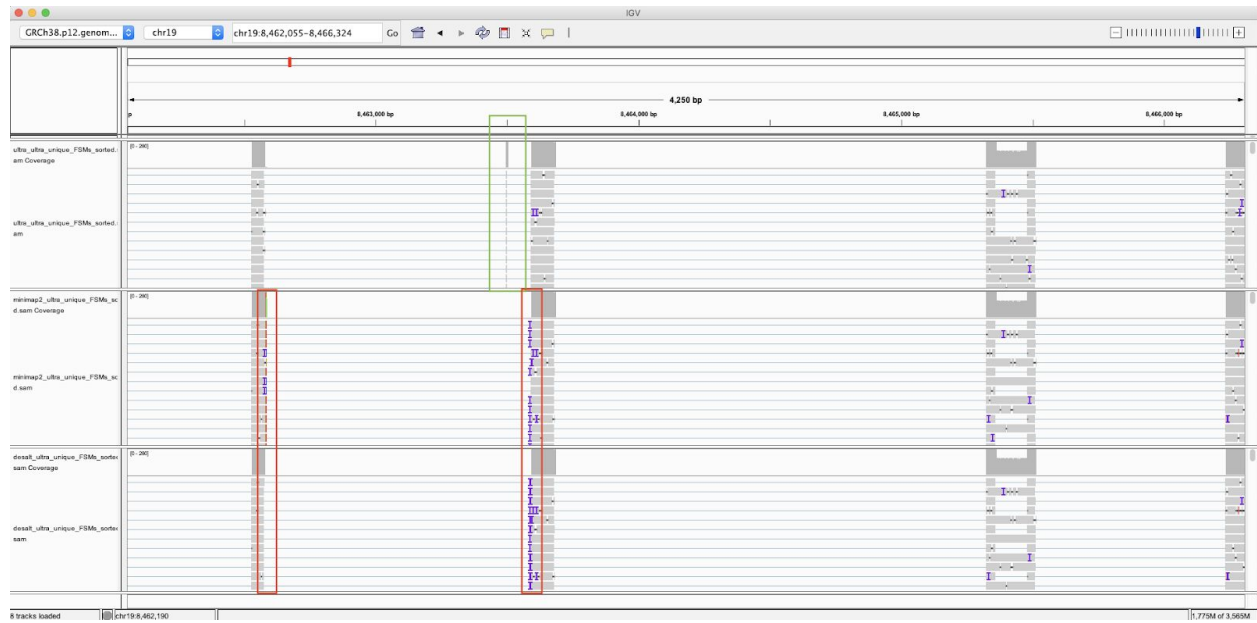
**Figure S8.** Histogram of the number of distinct isoform structures (y-axis) for which uLTRA aligned FSM reads to, but deSALT\_GTF and minimap2\_GTF did not. The x-axis shows the FSM read support for these isoforms. Panel (A) shows the DROS dataset and (B) the ALZ dataset.



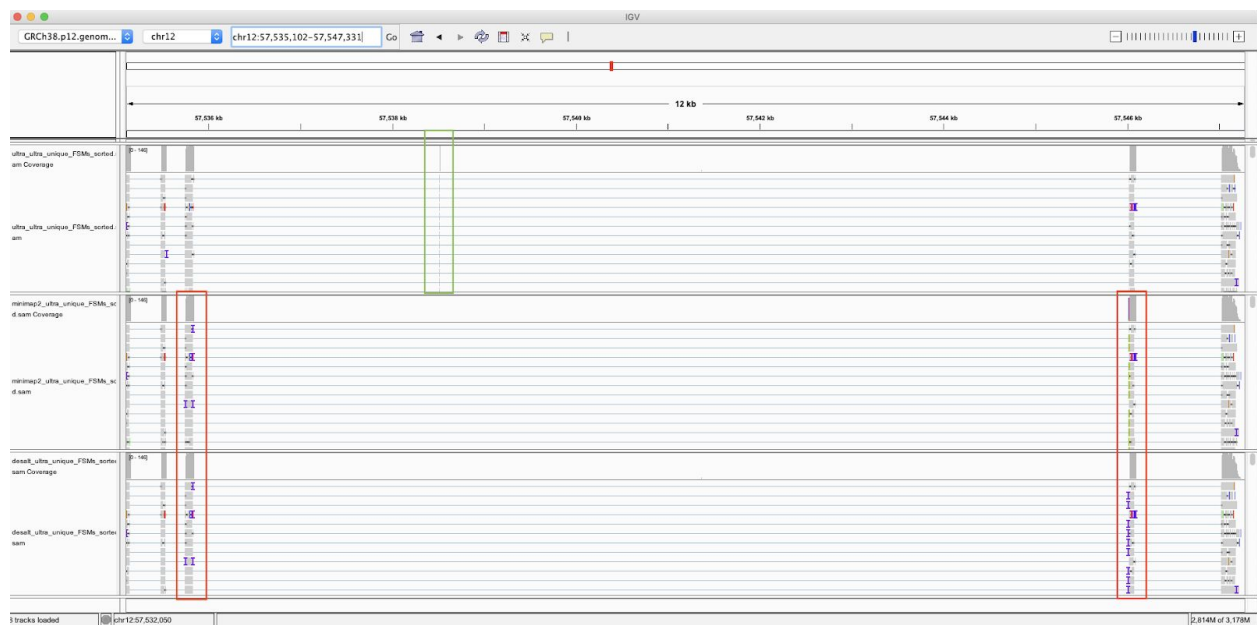
A



B

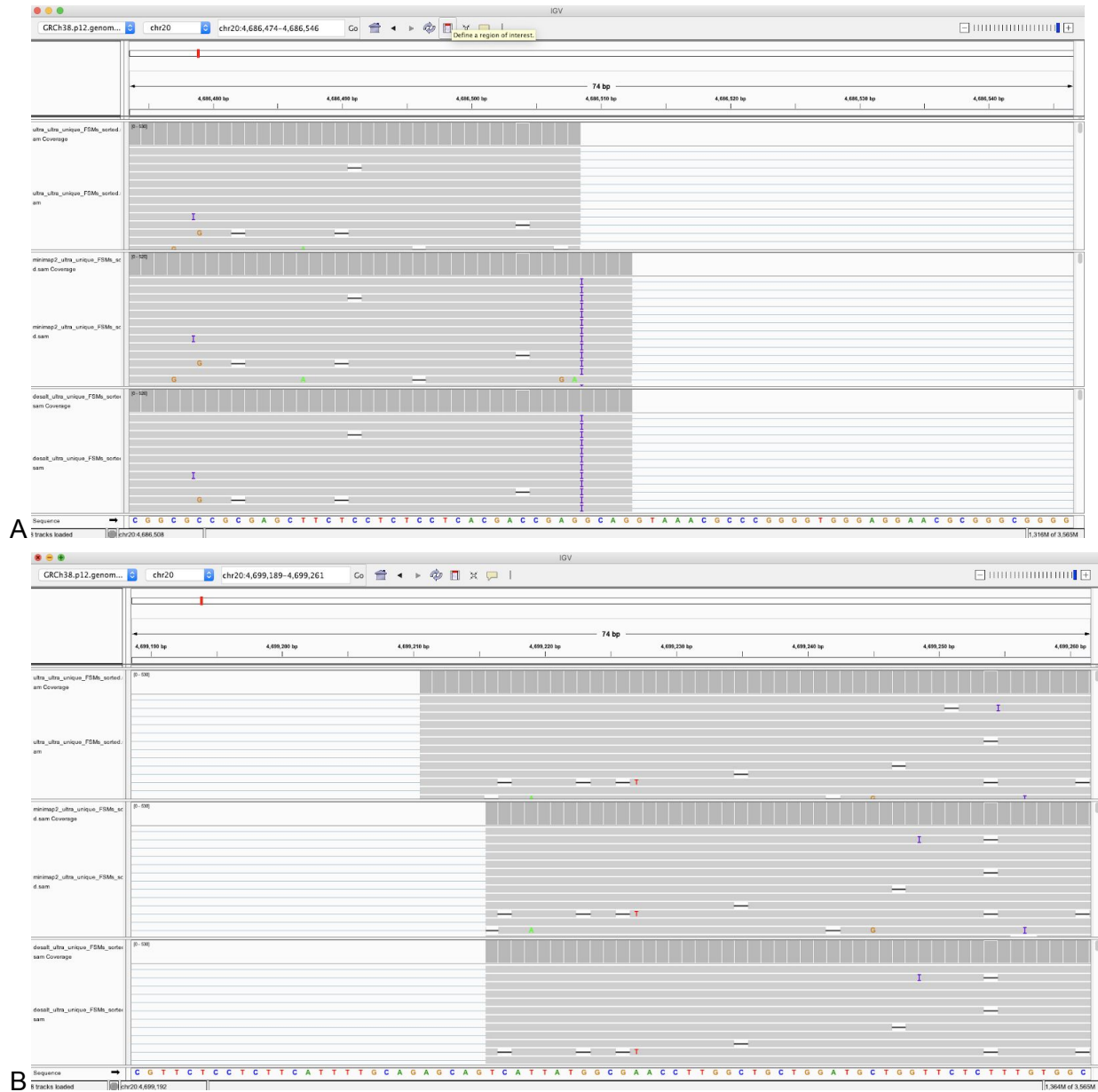


C

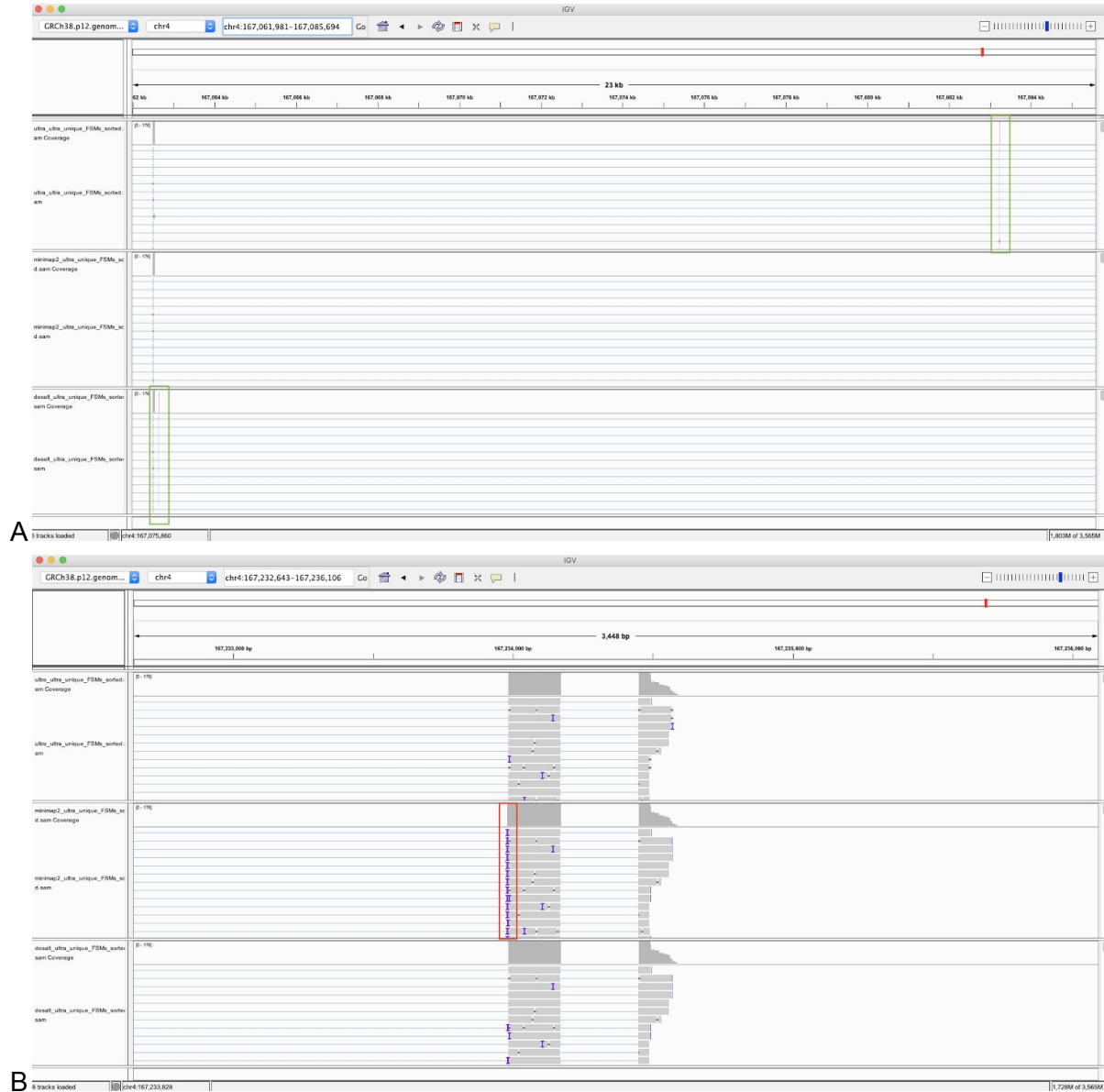


D

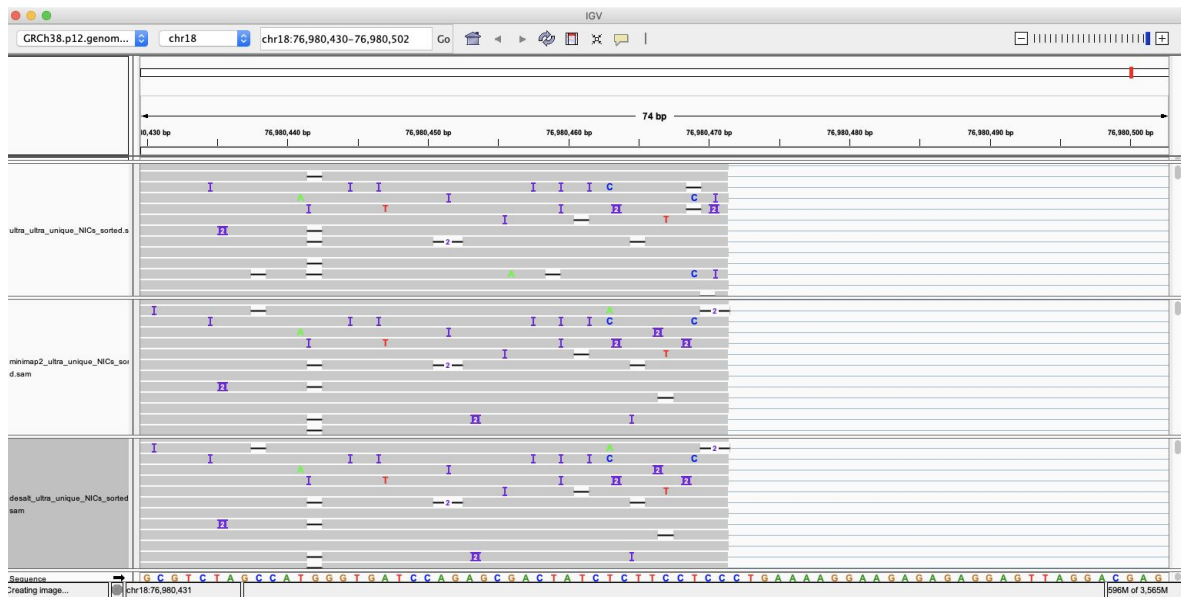
**Figure S9.** IGV tracks of FSM splice read alignments to unique FSM isoforms found by uLTRA (first track) by aligning to small exons (6-8nt), with corresponding aligned reads of minimap2 (second track) and deSALT (third track). The green box highlights the best fit exon alignment. The red box highlights misalignments around the junctions caused by the unaligned small exons. (A) 920 reads mapping to transcript ENST0000292807.9 (AP2 gene). (B) 770 reads mapping to transcript ENST0000609360.6 (APBB gene). (C) 167 reads mapping to ENST0000325495.9 (HNRNPM gene). (D) 136 reads mapping to ENST0000543672.5 (DCTN gene).



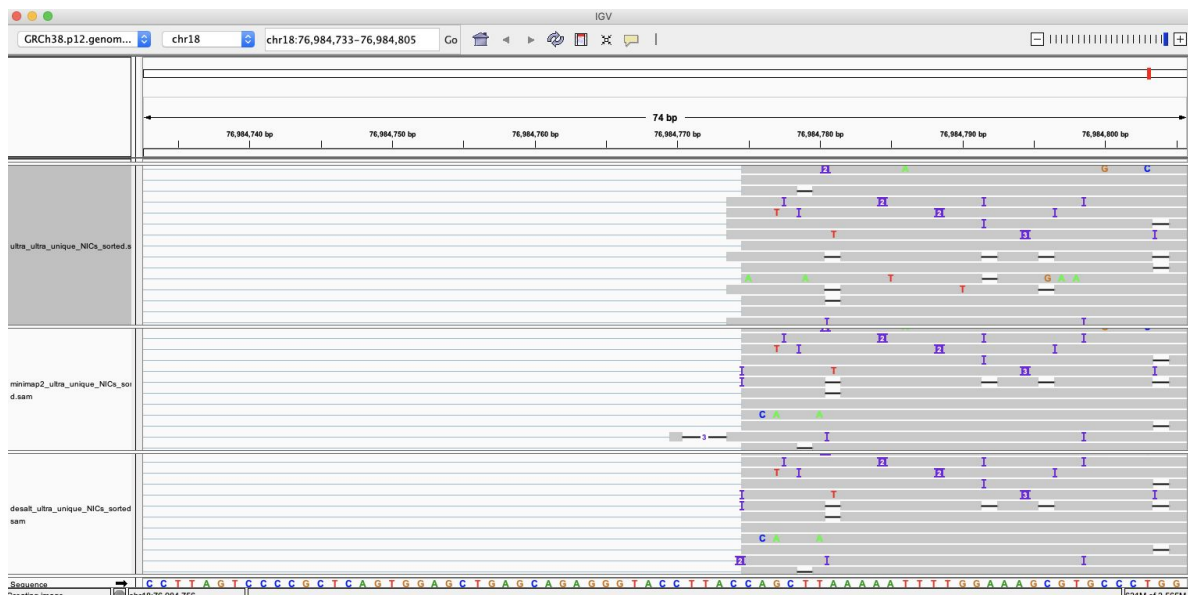
**Figure S10.** Potential subtle misalignment of 530 reads around junction for minimap2 (second track) and deSALT (third track) cause structural change in splice junction. Both minimap2 and deSALT place a 5nt segment in the upstream junction (with a 1nt insertion) instead of placing the 5nt segment (full match) in the downstream junction (B). This is caused by overfitting the alignment to match a GT-AG junction with junction specific penalties. In this gene, it appears as a small variant between the sample and the reference. Note that since the correct annotation is unknown for biological data, hence this may be a correct variation. However, uLTRA achieves higher identity in its alignment over the junction.



**Figure S11.** Discordant alignments of 176 reads between uLTRA (top track), minimap2 (middle track) and deSALT\_GTF (bottom track) to the SPOCK gene. For these reads, uLTRA and deSALT\_GTF align a 9nt portion of the reads to two different exons (A) while minimap2\_GTF does not align this region and is instead present as an insertion in the downstream exon (B). uLTRA chooses the upstream exon because of the deterministic implementation of taking the closest segment to the downstream hit in the traceback vector of the collinear MAM-chaining solution.



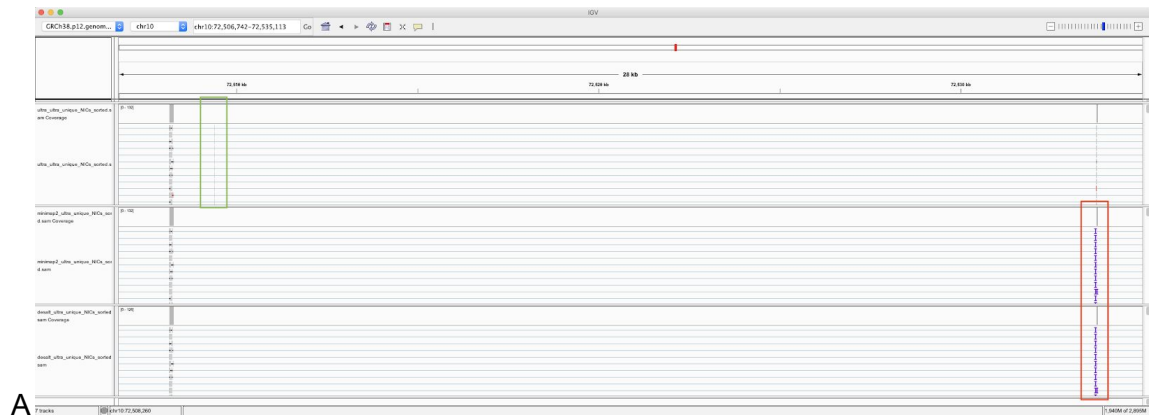
A



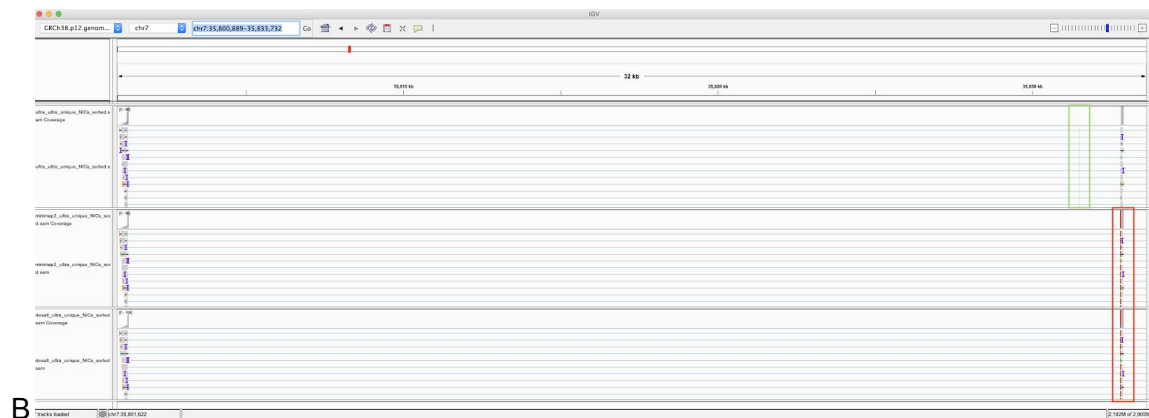
B

**Figure S12.** IGV tracks of splice read alignments to a junction of the MBP gene. Homopolymer stretches of C both upstream (A) and downstream (B) of the junction creates ambiguity in alignment. deSALT\_GTF and minimap2\_GTF always align to the CT-AC junction by creating insertions of C at downstream junctions if needed, while uLTRA chooses a CT-TA junction for the reads with a homopolymer stretch of four Cs. The CT-AC matches an FSM transcript while the CT-TA junction creates a NIC transcript (1208 reads).

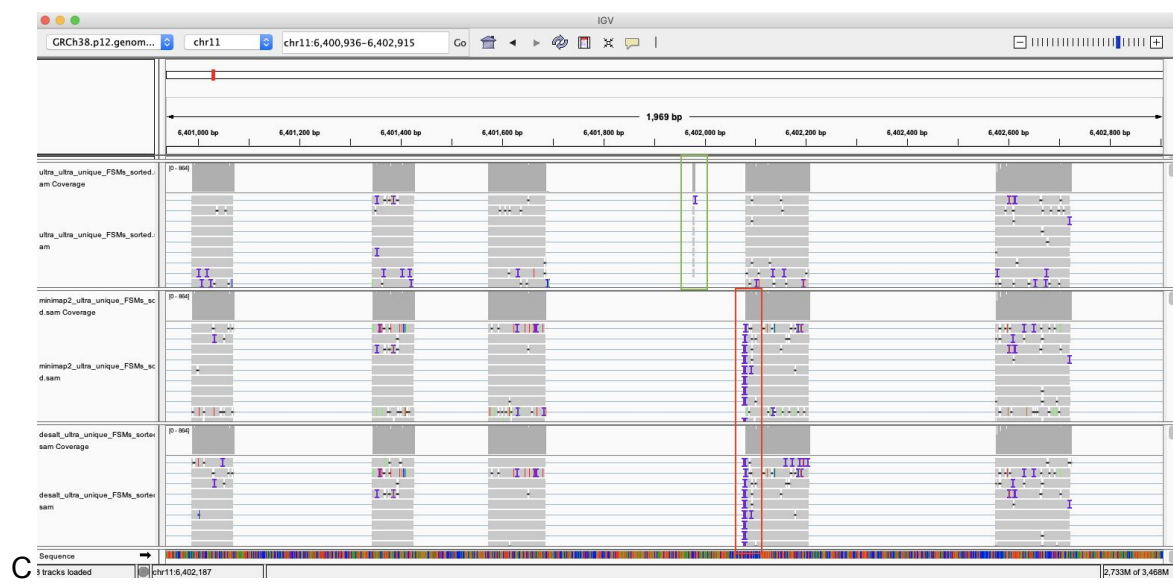




A



B



C

**Figure S13.** IGV tracks of splice read alignments to unique NIC isoforms found by uLTRA (first track) by aligning to small exons (5–6nt), with corresponding aligned reads of minimap2 (second track) and deSALT (third track). A green rectangle highlights the best fit exon alignments. The red box highlights misalignments around the junctions caused by the unaligned small exons. (A) 126 reads mapping to a NIC transcript (with a 6nt exon) from the MICU1 gene. (B) 81 reads mapping to a NIC (with a 5nt exon) transcript of the SEPTIN7 gene. (C) 27 reads mapping to a NIC transcript (with a 6nt exon) from the APBB1 gene.

## References

- Daily, Jeff. 2016. "Parasail: SIMD C Library for Global, Semi-Global, and Local Pairwise Sequence Alignments." *BMC Bioinformatics* 17 (February): 81.
- Fernandes, Francisco, and Ana T. Freitas. 2014. "slamEM: Efficient Retrieval of Maximal Exact Matches Using a Sampled LCP Array." *Bioinformatics* 30 (4): 464–71.
- Li, Heng. 2018. "Minimap2: Pairwise Alignment for Nucleotide Sequences." *Bioinformatics* 34 (18): 3094–3100.
- Liu, Bo, Yadong Liu, Junyi Li, Hongzhe Guo, Tianyi Zang, and Yadong Wang. 2019. "deSALT: Fast and Accurate Long Transcriptomic Read Alignment with de Bruijn Graph-Based Index." *Genome Biology* 20 (1): 274.
- Li, Yu, Renmin Han, Chongwei Bi, Mo Li, Sheng Wang, and Xin Gao. 2018. "DeepSimulator: A Deep Simulator for Nanopore Sequencing." *Bioinformatics* 34 (17): 2899–2908.
- Sahlin, Kristoffer, Botond Sipos, Phillip L. James, Daniel J. Turner, and Paul Medvedev. n.d. "Error Correction Enables Use of Oxford Nanopore Technology for Reference-Free Transcriptome Analysis." <https://doi.org/10.1101/2020.01.07.897512>.
- Faucon, P. C., Balachandran, P., and Crook, S. "SNaReSim: Synthetic Nanopore Read Simulator," *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, Park City, UT, 2017, pp. 338-344, doi: 10.1109/ICHI.2017.98.
- Šošić, Martin, and Mile Šikic. 2017. "Edlib: A C/C++ Library for Fast, Exact Sequence Alignment Using Edit Distance." *Bioinformatics* 33 (9): 1394–95.
- Stöcker, Bianca K., Johannes Köster, and Sven Rahmann. 2016. "SimLoRD: Simulation of Long Read Data." *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btw286>.
- Yang, Chen, Justin Chu, René L. Warren, and Inanç Birol. 2017. "NanoSim: Nanopore Sequence Read Simulator Based on Statistical Characterization." *GigaScience* 6 (4): 1–6.