Supplemental Materials for:

Transcriptional fates of human-specific segmental duplications in brain

Max L. Dougherty, Jason G. Underwood, Bradley J. Nelson, Elizabeth Tseng, Katherine M. Munson, Osnat Penn, Tomasz J. Nowakowski, Alex A. Pollen, and Evan E. Eichler

Table of Contents

Description of other Supplemental Files

Supplemental_Tables.xlsx
- Table S1. Probe design, including targeted gene families and number of probes
- Table S2a. HSD1 probe design sequences
- Table S2b. HSD2 probe design sequences
- Table S3. Barcoded end sequences
- Table S4. SMRT cells sequenced
- Table S5. Summary sequence statistics
- Table S6. Analyzed gene set
- Table S7. Fusions vs. truncations in 3' truncated genes
- Table S8. Quantification of isoform abundance for Tigger-derived exon

Novel Isoforms FASTA file: Supplemental_File_S1.fasta
Novel Isoforms GFF file: Supplemental_File_S2.gff
Custom Iso-Seq processing script: Supplemental_File_S3.py
- See also: https://github.com/EichlerLab/isoseq_pipeline

Supplemental Methods

cDNA synthesis oligonucleotides: We synthesized specialized poly(dT) oligonucleotides to prime first-strand cDNA synthesis (Integrated DNA Technologies [IDT]) with the following configuration: 5'-AAGCAGTGGTATCAACGCAGAGT(BC16bp)TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN-3', where the sequence BC16bp encodes one of 96 of the 16 bp barcodes, V=(A,G,C), and N=(any base). We synthesized oligonucleotides for second-strand synthesis with the following configuration: 5'-AAGCAGTGGTATCAACGCAGAGT(BC16bp)ATACGATTTAGGTGACACTATAGG-3' where the sequence BC16bp encodes one of 96 of the 16 bp barcodes. The template switch oligonucleotide, a chimeric RNA-DNA sequence, was synthesized: AAGCAGTGGTATCAACGCAGAGTACATrGrGrG
For cDNA amplification, the 5' flanking sequence in the first- and second-strand oligonucleotides was utilized for PCR: /5Phos/AAGCAGTGGTATCAACGCAGAGT.

Mapping of putative full-length cDNA: To take advantage of the higher read qualities available through the Iso-Seq "clustering" pipeline (in which multiple reads are used to generate a consensus isoform), but to avoid the potential for confounding by reads from separate paralogs being merged in the same cluster, we performed this step in partitioned genomic regions (as described in [Kronenberg et al. 2018]). Regions are split to ensure that no pair of SD "mates" is found in any one region. Only confidently mapped reads (MAPQ > 40) are input into the clustering step, and clustering is performed separately in each region, generating consensus isoforms without contamination from other paralogs (Figure S6).

We further performed the "collapse" step whereby consensus isoforms are mapped and the mappings are used to remove redundant isoforms. In practice, we found that this generated a greater number of "nonredundant isoforms" for most paralogs than would be expected and that many appeared to be fragments of isoforms, especially for genes with longer transcripts for which we were less likely to have captured full-length isoforms on single reads. Therefore, it was necessary to merge more than one

fragment isoform from the final output of this modified Iso-Seq pipeline to generate a full-length gene model.

Finally, newly determined isoforms were assessed for support by other data sources. This includes reads of 5' ends of RNA molecules generated with cap analysis of gene expression (CAGE) data from the FANTOM5 consortium (Lizio et al. 2015), 3' ends of polyadenylated RNA molecules (poly(A)-seq) from Leslie, Mayr, and colleagues (Lianoglou et al. 2013) remapped to GRCh38, and various PacBio RNA-seq datasets, including: H1 human embryonic stem cell (GEO accession GSM1254204), Iso-Seq whole transcriptome from human brain with Alzheimer's disease (https://downloads.pacbcloud.com/public/dataset/Alzheimer_IsoSeq_2016/ accessed Dec 2016), whole transcriptome from human brain, liver, and heart (http://datasets.pacb.com.s3.amazonaws.com/2014/Iso-seq_Human_Tissues/list.html accessed Dec 2016), and Iso-Seq whole transcriptome generated from MCF-7 human breast cancer cell line (https://github.com/PacificBiosciences/DevNet/wiki/IsoSeq-Human-MCF7-Transcriptome accessed Dec 2016).

Secondary analysis of long-read RNA-seq data: *Proportion of fusion (duplication-spanning) versus truncation reads:* For each duplicate gene investigated, a constitutive exon close to the breakpoint was chosen as an anchor point, and mapped pFL reads (MAPQ > 40) including that exon were selected in order to mitigate non-full-length reads from inflating the "truncated" count. Counts of reads that contain spliced exons beyond the duplication breakpoint were calculated using BEDTools (Quinlan and Hall 2010). Based on this proportion, a duplicate gene was designated as "primarily truncation" (<0.2), "primarily fusion" (>0.8), or "both" (0.2–0.8).

Splicing disorder: The number of supporting pFL reads for *FCGR1A* and *FCGR1B* isoforms output by the "collapse" step of the modified Iso-Seq pipeline (minimum 2 pFL reads) were used as an approximation of relative isoform abundance. Shannon's entropy was calculated using the isoform

abundances for each paralog with the entropy package (v1.2.1) (Hausser and Strimmer 2009, *J. Mach. Learn. Res.* 10, 1469-1484).

Probes for *in situ* hybridization: Probes used for RNA *in situ* hybridization were synthesized within a pCI-NEO vector flanked by XhoI-NotI cloning sites to the antisense strand of target mRNAs (Promega): probe ARHGAP11AB-e1e2e3e4 ("AB") was designed to bases 146 to 805 of NM_001039841.1; probe ARHGAP11A-e12-utr ("A") to bases 3286 to 4005 of NM_014783.4; probe ARHGAP11B-07utr ("B") to bases 2983 to 3688 of NR_038253.1. Note that probe "B" is not homologous to any transcribed part of the ancestral *OTUD7A* gene, and thus is not predicted to detect *OTUD7A* expression.

Supplemental Figures

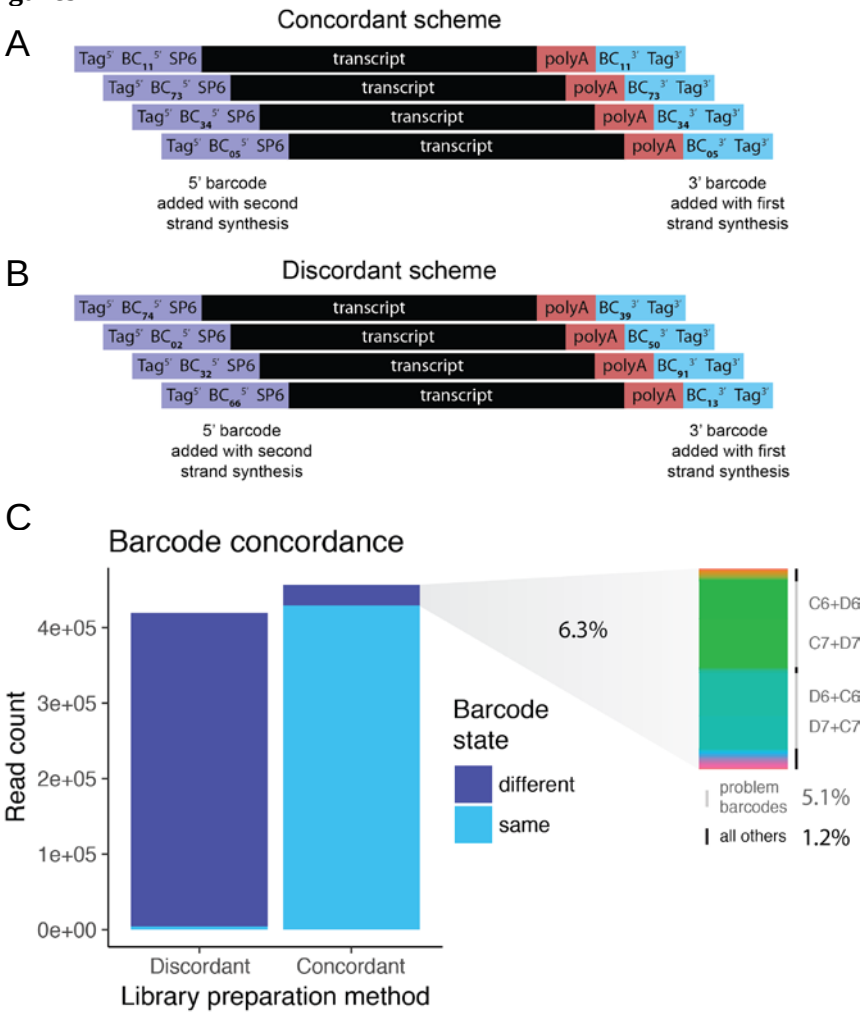


Figure S1. Barcoding strategy. A) In the concordant barcode scheme, 1 of 96 identical barcodes is appended to the 3' and 5' ends of transcript sequences during first strand and second strand synthesis, respectively. This allows for detection of false chimeric sequences generated during amplification. **B)** In the discordant barcode scheme, 1 of 96 barcodes is appended at random to 3' and 5' ends of transcript sequences. In this method, the combination (96x96) can be used as a pseudo-unique molecular identifier to measure the frequency of PCR duplicates. **C)** The concordance of reads from libraries prepared under both schemes is shown, measured as number of putative full-length (pFL) reads with barcodes that are the same or different. Among discordant-scheme libraries, we obtain 1.0% reads with matching barcodes, consistent with a 1 in 96 chance. Among concordant-scheme libraries, we obtain 93.7% reads with matching barcodes, indicating that 6.3% are not as expected. When we examine which barcode pairs comprise these unexpectedly discordant reads, we observe that the majority (80%) are derived from two particular pairings of barcodes (in both orientations), here labelled by their position in a 96-well plate. Contamination specifically between the barcodes of wells C6 and D6, and wells C7 and D7, are mostly responsible. Because the pattern holds true for all concordant-scheme libraries over multiple separate library preparations, we believe this is most likely due to a manufacturing error that led to contamination. Eliminating those “problem pairings”, we observe a rate of 1.2% discordancy, indicating that chimeric formation during PCR is rare.

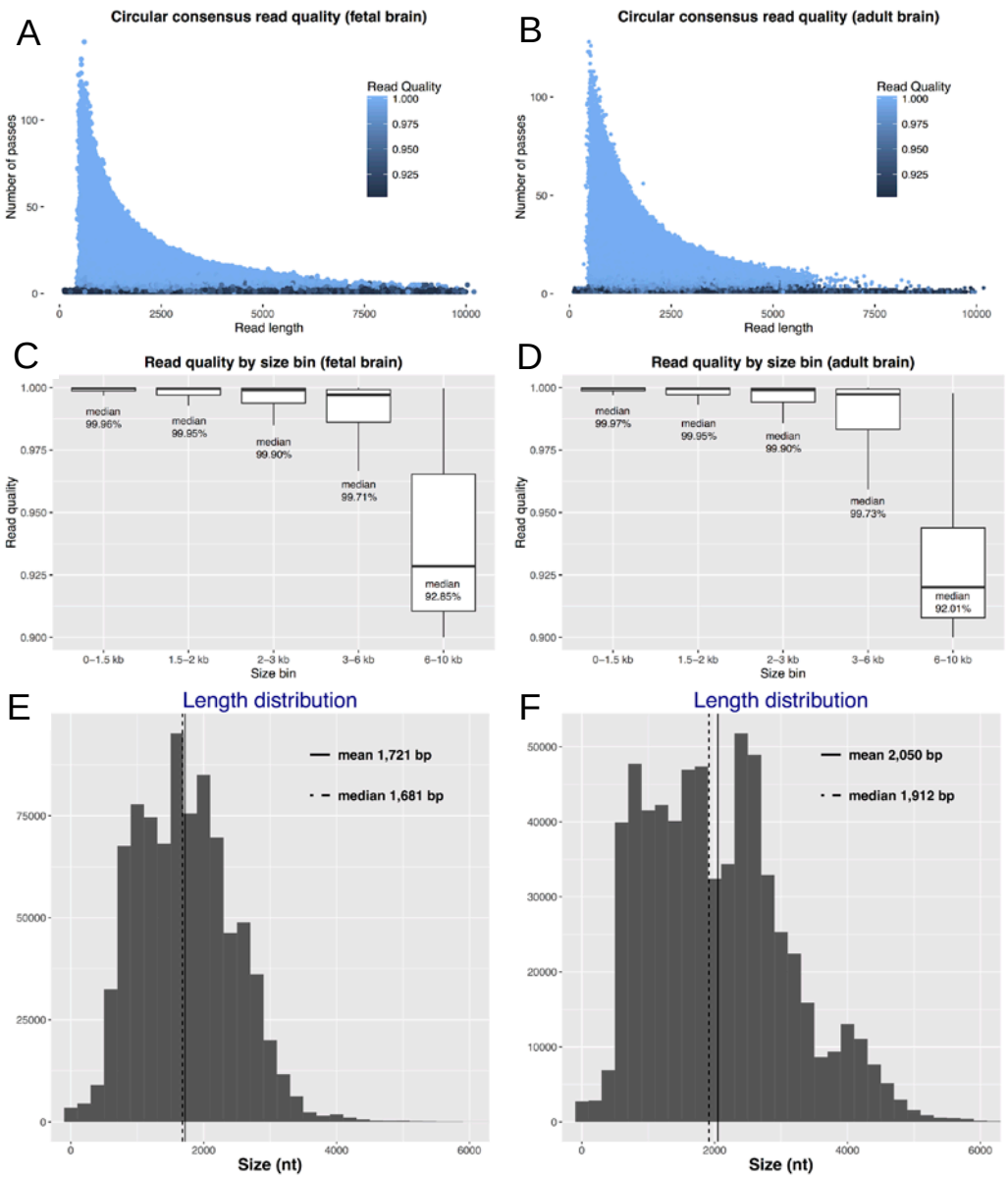


Figure S2. CCS read quality is highly dependent on read length. In both fetal brain (**A**) and adult brain (**B**), there is an inverse relationship between read length and the number of passes by the sequencing polymerase. This determines the number of subreads from which a given read is generated, which directly affects read quality. Each point represents a CCS read. Reads with a high number of passes (i.e., subreads) have very high average read quality. When reads are binned by size (**C**, **D**), we see that reads smaller than 3 kbp have on average >99.9% read quality, but for reads over 6 kbp that falls to ~92%. The ICE algorithm clusters these low-quality reads and uses the resulting multiple sequence alignment to arrive at cluster-based isoforms with >99% read quality. Data is shown for HSD1-enriched cDNA, for which we generated the most sequencing reads. Quality values are taken from the "rq" field of post-CCS bam files. Size distributions of pFL reads designated for (**E**) adult brain and (**F**) fetal brain.

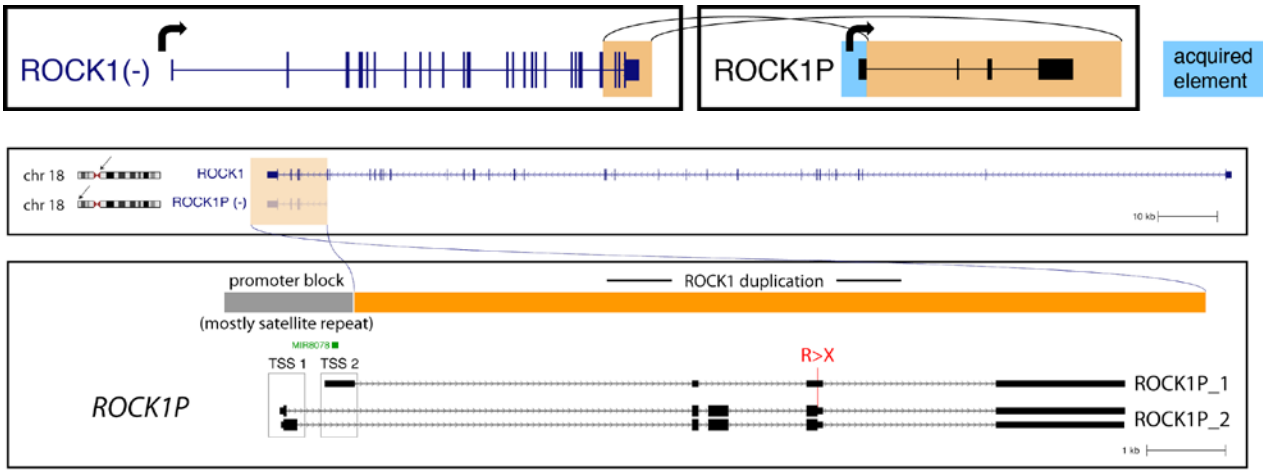


Figure S3. Promoter exaptation in the 5' truncated duplicate gene *ROCK1P*. The final four exons and preceding splice donor of the large gene *ROCK1* underwent an inverted intrachromosomal duplication on Chromosome 18, generating *ROCK1P*. "Rescue" of the 5' truncated gene is facilitated by promoter borrowing from the adjacent duplication block, which creates two nearby transcription start sites (TSS 1 and TSS 2) that overlap a microRNA identified in testis. *ROCK1P* has highest expression in the testis, in contrast to the ancestral widely expressed *ROCK1*, likely indicating a role for this acquired promoter in *ROCK1P*'s expression pattern. A stop gain on the penultimate exon of *ROCK1P* makes the open-reading frame (ORF) even shorter than by duplication alone, possibly indicating neutrality of the coding sequence.

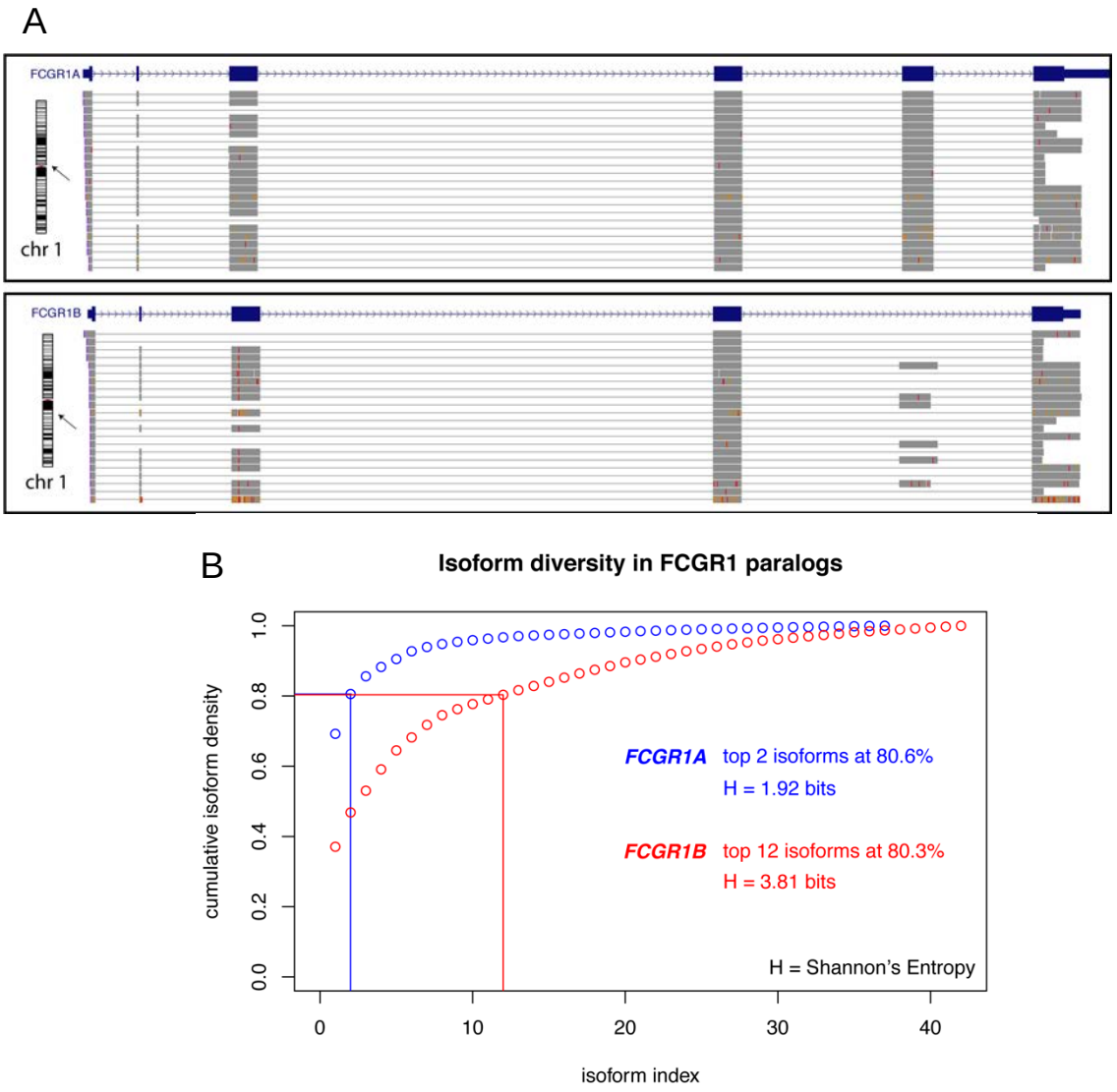
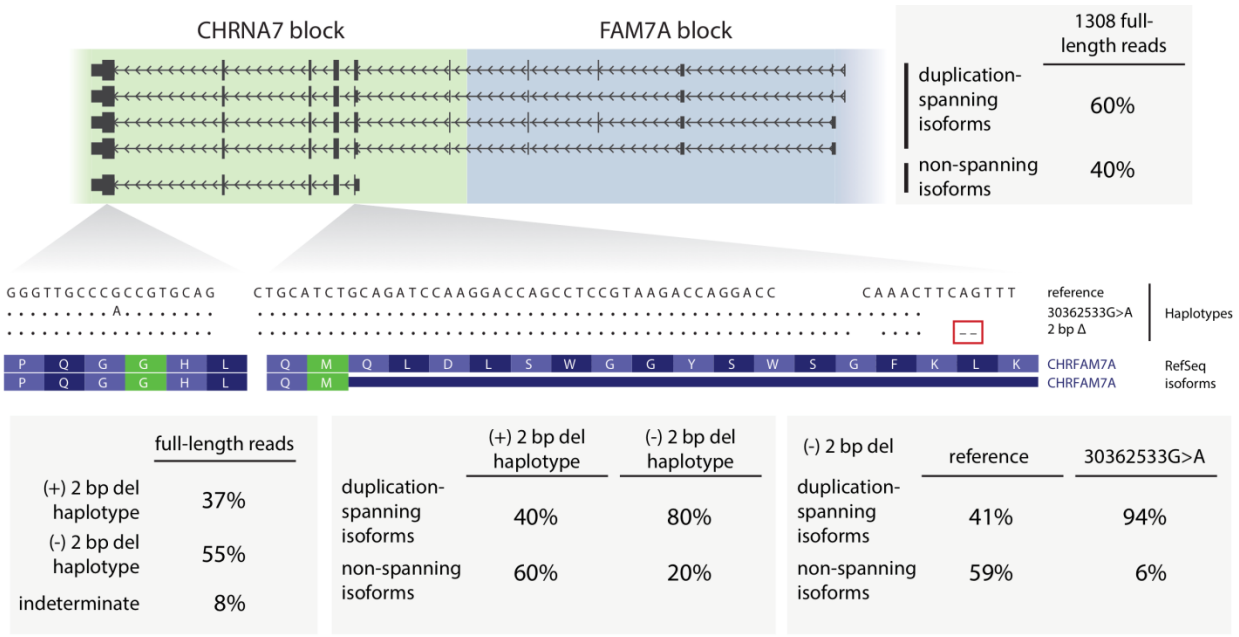


Figure S4. Relaxed selection on splicing results in increased disorder of isoforms. A) A random subset of ~25 reads each for *FCGR1A* and *FCGR1B* is shown. A visually striking difference in the orderliness of isoforms is apparent, with *FCGR1B* appearing substantially more disorderly. When compared to the ancestral paralog *FCGR1A*, *FCGR1B* displays a greater diversity of splice isoforms, particularly due to variable choice of splice donor and acceptor sites at exon 5, as well as exclusion of exon 5. In *FCGR1A*, this lack of incorporation of this exon is not observed (<0.4%). **B)** Cumulative distributions of isoform abundances in *FCGR1A* (blue) and *FCGR1B* (red) are plotted. The two most abundant *FCGR1A* isoforms comprise 80.6% of the total *FCGR1A* reads, while for *FCGR1B*, it is not until the 12^{th} most abundant isoform that 80.3% of total reads is reached, indicating a significant "flattening out" of relative isoform abundances ($p = 1.3 \times 10^{-7}$), also manifesting as an increase in the entropy of this distribution.



	full-length reads
(+) 2 bp del haplotype	37%
(-) 2 bp del haplotype	55%
indeterminate	8%

	(+) 2 bp del haplotype	(-) 2 bp del haplotype
duplication-spanning isoforms	40%	80%
non-spanning isoforms	60%	20%

(-) 2 bp del	reference	30362533G>A
duplication-spanning isoforms	41%	94%
non-spanning isoforms	59%	6%

Figure S5. Identification of 5' truncated *CHRFAM7A* isoform. SMRT sequencing of probe-enriched cDNA generated from pooled fetal and adult brain yielded 1,308 pFL *CHRFAM7A* reads (MAPQ > 40). Of these reads, 60% spanned the *CHRNA7* duplication boundary and included upstream exons from the *FAM7A* block, while 40% of reads were initiated from a transcription start site (TSS) internal to the *CHRNA7* block, in the vicinity of the splice acceptor of the exon paralogous to exon 6 of canonical *CHRNA7* (referred to as *CHRFAM7A* exon 6). While we do not have access to the genotypes of the individuals from whom these reads were generated, three alleles can be discerned from variants within the transcribed sequence: the reference allele, the 30363533G>A allele, and the 2 bp Δ allele. The 2 bp Δ allele causes a frameshift in the ORF found in some of the duplication-spanning isoforms but is outside the ORF that begins in *CHRFAM7A* exon 6, which is contained in the other duplication spanning isoforms as well as the non-spanning isoform. The 2 bp Δ transcripts have a smaller proportion of duplication spanning isoforms when compared to transcripts without the deletion. Reads with the 30363533G>A allele are almost exclusively duplication-spanning, while the other two alleles share a similar ratio. Naively, the non-spanning isoform should have a greater probability of having the relevant, translated ORF, given it is unaffected by the 2 bp Δ and has a 5' UTR that is shorter and composed of fewer exons.

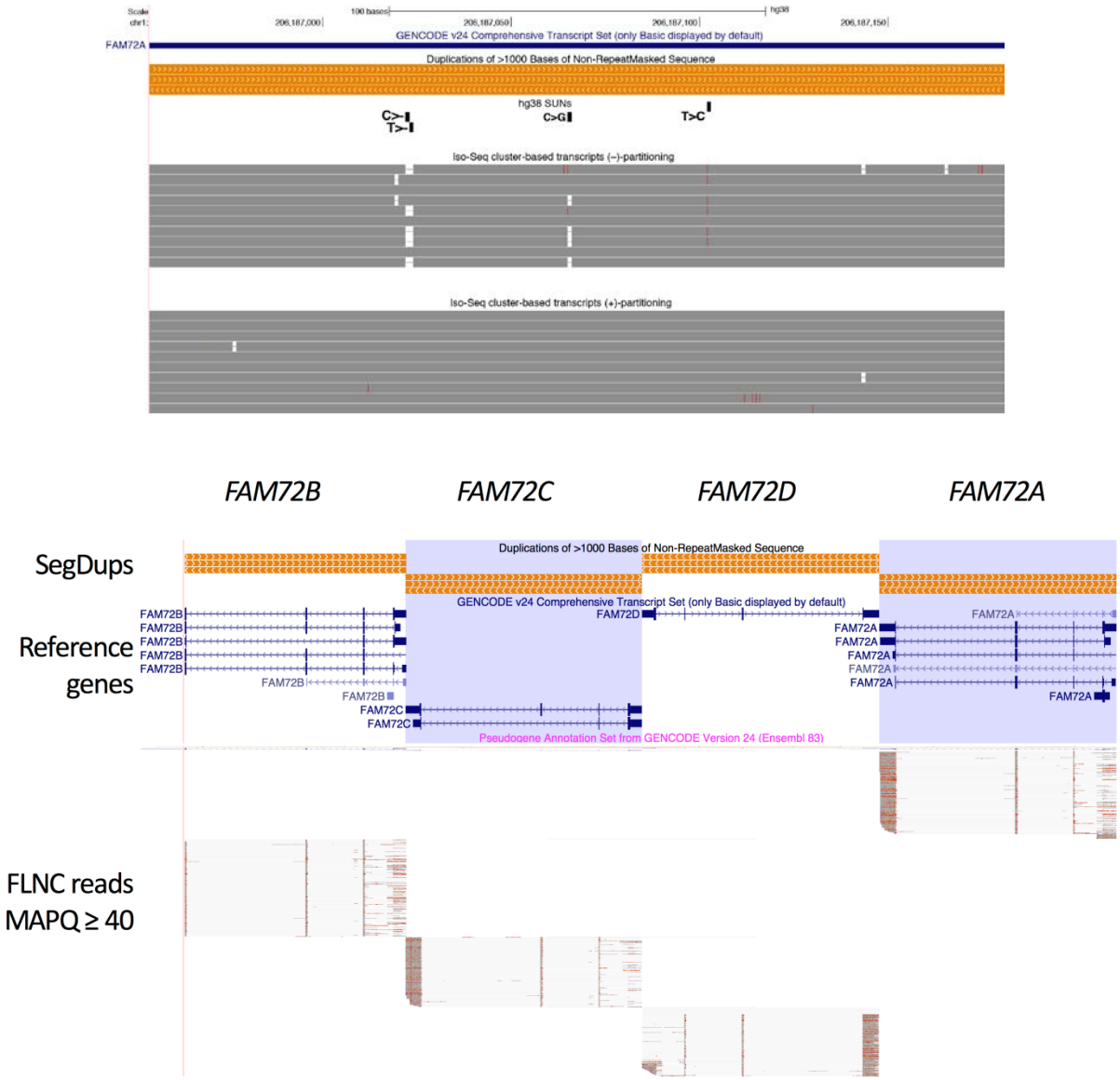


Figure S6. Pre-partitioning of reads mitigates generation of *in silico* isoform chimeras. Shown is a UCSC Genome Browser screenshot of the 3'-UTR of *FAM72A*, a gene with four reference paralogs. A cluster of paralogous sequence variants (PSVs) that distinguish this paralog from all others are shown below the segmental duplication track (orange bars, indicating >99% sequence identity). Below the PSVs are two tracks showing a random selection of 10 isoforms from the output of ICE mapped back to the genome using GMAP with a mapping quality cutoff of 40. The above track uses the default application of the ICE pipeline, and *in silico* chimeric reads are apparent, as a consequence of the clustering and alignment of reads from not only *FAM72A* but also other *FAM72* paralogs, resulting in SNVs corresponding with known PSVs. The bottom track uses ICE after an additional "partitioning" step, wherein pFL reads are first mapped with a mapping quality cutoff of 40, then partitioned into separate regions to ensure that reads belonging to different paralogs are not clustered together during ICE. Applying this strategy, the spurious variants disappear, indicating the output of ICE is no longer confounded by mixing paralogs.

Supplemental Note

While the most notable or novel features of isoforms detected from targeted HSD gene families are already discussed (main text/supplement), we also present a short summary by gene family with reference to specific and novel isoforms that were discovered during this analysis (named in isoform supplemental files). See also Figure 3 and Table S6 for an overview of features of isoforms identified by gene family.

NOTCH2NL: We characterize a number of isoforms for the truncated *NOTCH2NL* duplicated copies, which differ mainly by two alternative splicing events: 1) use of an alternate splice acceptor in exon 2, which shortens the exon by 8 bp, and 2) retention of the final intron. These include *NOTCH2NL_1*, *NOTCH2NL_2*, *NOTCH2NL_3*, *NOTCH2NL_4*, *NOTCH2NLB_1*, *NOTCH2NLC_1*, *NOTCH2NLD_1*, and *NOTCH2NLD_2*. We also note frequent fusion transcription between *NOTCH2NL* paralogs and downstream *NBPF* genes.

SRGAP2: In addition to previously described *SRGAP2* isoforms, we identify isoform *SRGAP2B_2*, which contains the ORF-disrupting 61 bp exon (see Figure 6).

CD8B: We identify two isoforms for *CD8B2*, including the major isoform *CD8B2_1*, which is homologous to the major isoform of *CD8B*, and the minor isoform *CD8B2_2* whose final exon sits outside the duplication boundary, thus predicted to produce a different C-terminus (see also Figure 5).

ARHGAP11: We report a minor isoform of the ancestral gene *ARHGAP11A* (*ARHGAP11A_7*) that is an in-frame fusion with the downstream gene *SCG5*. We include three isoforms of *ARHGAP11B*, which demonstrate varying degrees of transcript fusion: *ARHGAP11B_3, ARHGAP11B_5,* and *ARHGAP11B_6* (see also Figure 4).

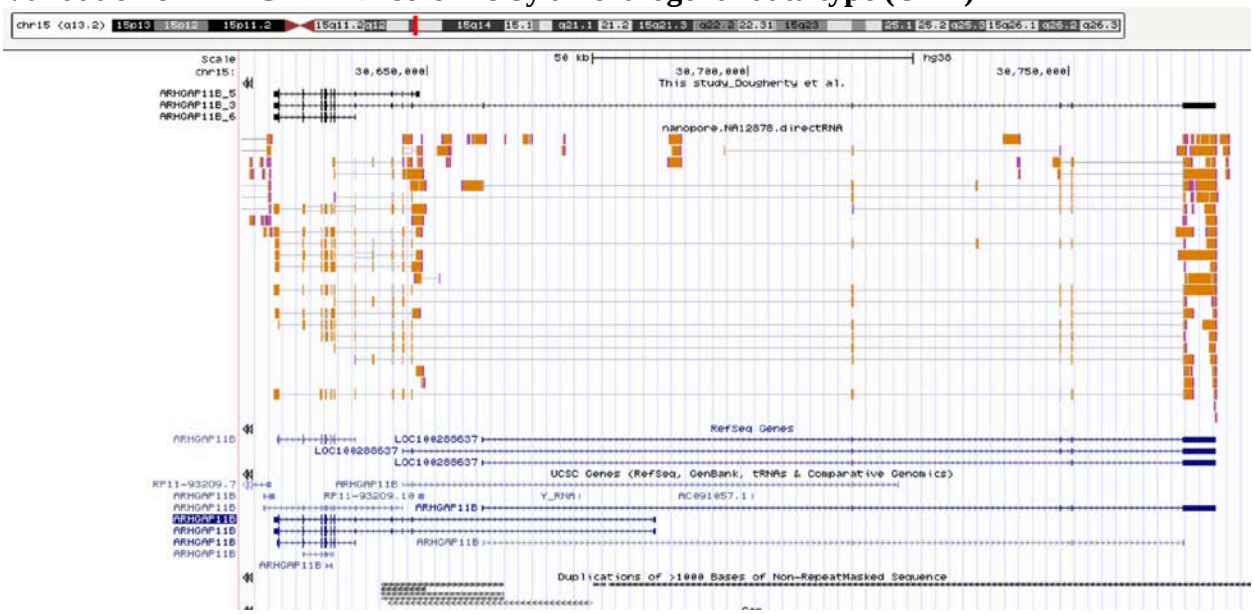
FMRPD and ***PTPN20*CP**: We report a truncated isoform of *FRMPD2B* (*FRMPD2B_1*) and two fusion isoforms of *PTPN20CP* (*PTPN20CP_1*, *PTPN20CP_2*).

CHRNA7: We include the truncated isoform of *CHRFAM7A* whose TSS is contained within the partial duplication of *CHRNA7* (*CHRFAM7A_1*, see also Figure S5).

ROCK1P: We report two isoforms of *ROCK1P1* (*ROCK1P1_1*, *ROCK1P1_2*), which derive their TSS from adjacent duplicate sequence (see also Figure S3).

GTF2I: We report two isoforms of the duplicate *GTF2IP1*, which appear to have a TSS found within the duplication and differ by retention of the final intron (*GTF2IP1_1*, *GTF2IP1_2*). Finally, for *GTF2I*, *GTF2IRD2*, and *GTF2IRD2B*, we include the isoforms detected that contained the alternative, Tigger7-derived, first exon (*GTF2I_1*, *GTF2IRD2_1*, *GTF2IRD2B_1*) as well as the noncoding fusion isoform between *GTF2IRD2* and *STAG3L2* (*GTF2IRD2_2*, see also Figure 7).

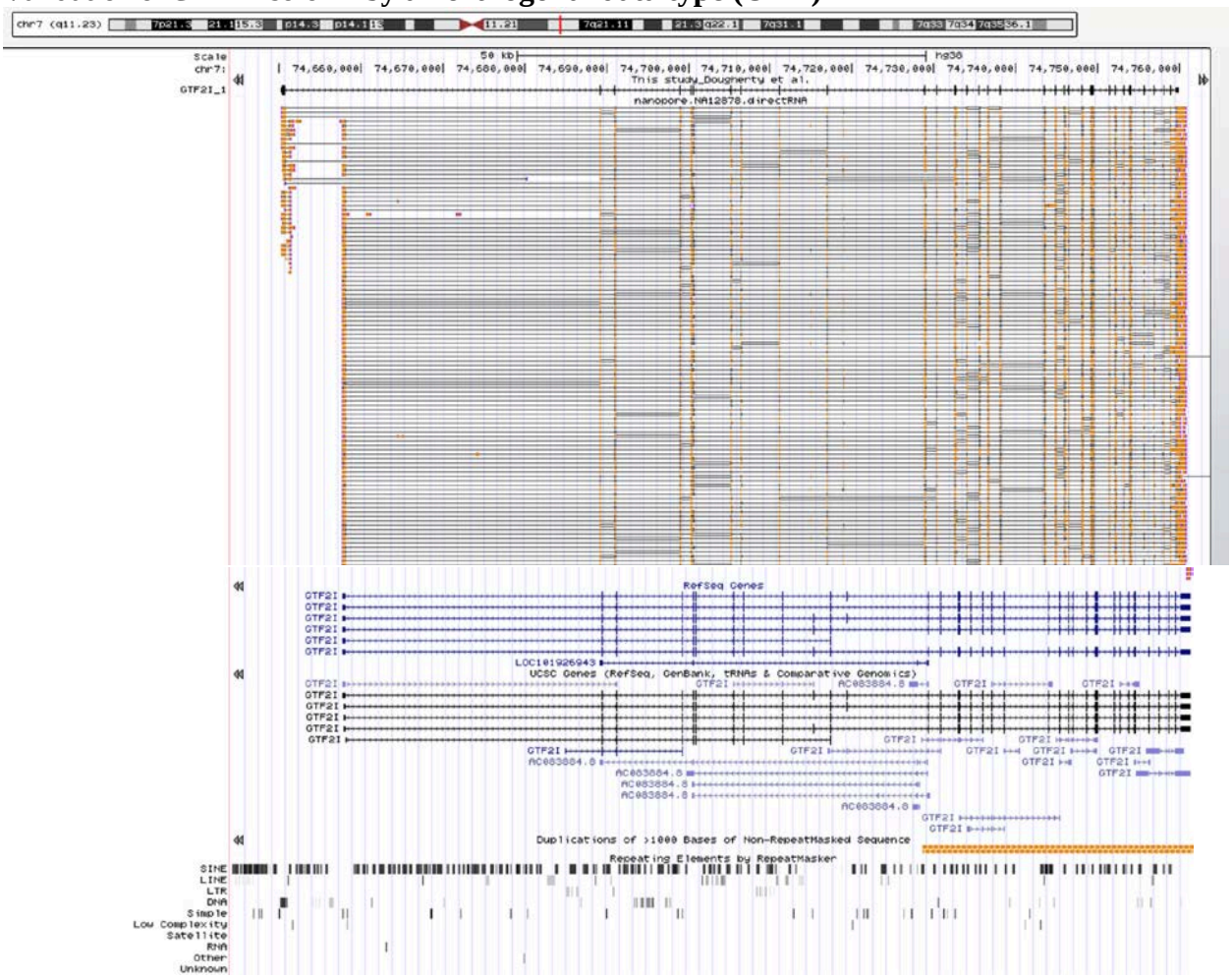
Validation of *ARHGAP11B* isoforms by an orthogonal data type (ONT)



UCSC Genome Browser snapshot for hg38 chr15:30,621,110-30,780,398 displaying (from top):
Black: ARGHAP11B isoforms described in this study
Orange: Oxford Nanopore Technologies (ONT) direct RNA-seq reads
Blue: RefSeq and UCSC Genes
Gray: Segmental duplication blocks (UCSC color code for 90-98% sequence identity)

Summary: **This study uncovered new isoforms of *ARHGAP11B* that contain extended 3' UTRs with respect to the RefSeq annotation.** ONT reads support the splicing architecture of longer isoforms that we observe. The ONT data does not detect any copies of the short form, *ARHGAP11B_6*, supporting our findings that it is less abundant than the longer isoforms. Note: RefSeq track currently has only the short form of *ARHGAP11B* on the left and then three different *LOC100288637* predicted transcripts. Our data and the ONT direct RNA-seq both support joining these *LOC100288637* exons to the *ARHGAP11B* gene body.

Validation of *GTF2I* isoform by an orthogonal data type (ONT)



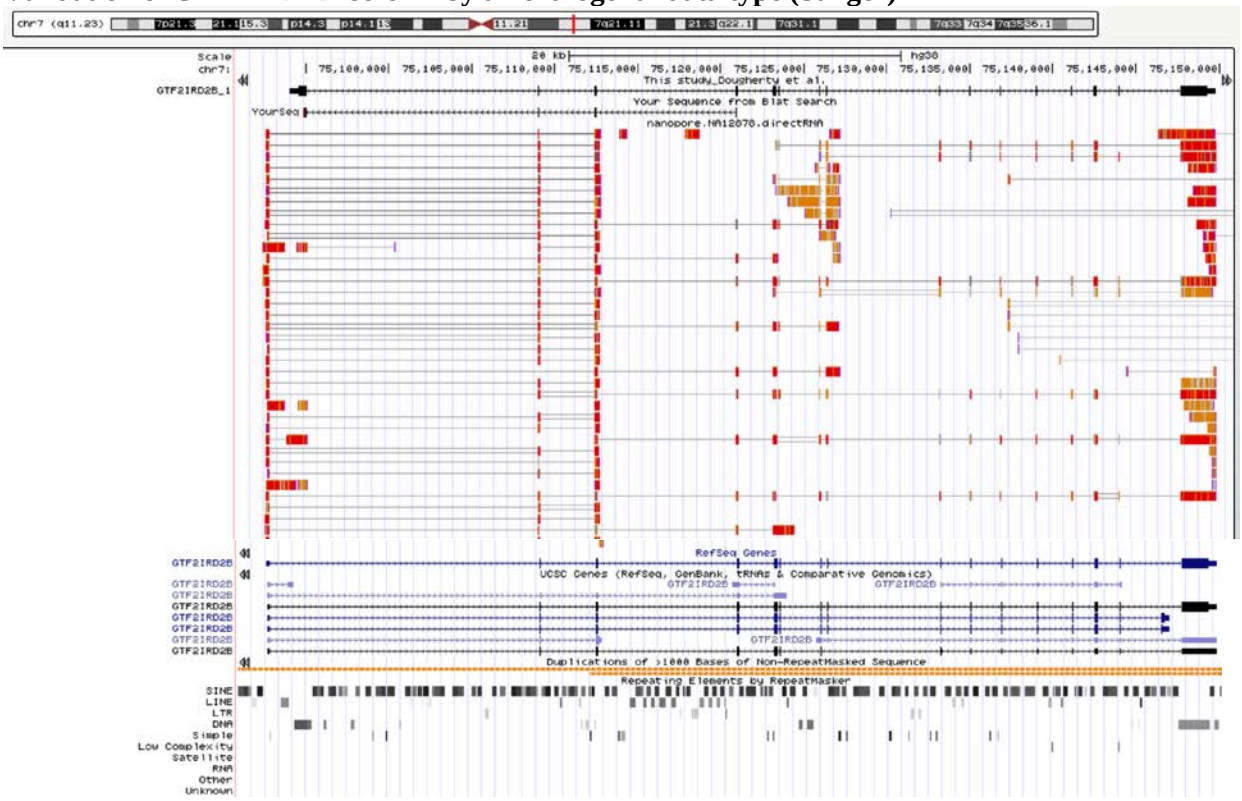
UCSC Genome Browser snapshot for hg38 chr7:74,644,295-74,763,681 displaying (from top):

Black: *GTF2I* isoform described in this study
Orange: Oxford Nanopore (ONT) direct RNA-seq reads
Blue: RefSeq and UCSC Genes
Orange: Segmental duplication blocks (UCSC color code for 98-99% sequence identity)
Black: Repeat elements by RepeatMasker

Summary: **This study uncovered a new isoform of the general transcription factor, *GTF2I*, with a new first exon encoded by a repeat element of the TcMar-Tigger 7 DNA family.** The ONT data are abundant in this region and a subset of the reads (~25%) are shown. ONT reads (n = 8) support the Tigger7 exon splicing to exon 2 of the canonical *GTF2I* isoform shown in RefSeq. In the RepeatMasker track, the repeat element at the coordinates of the first exon is:

Name: Tigger7
Family: TcMar-Tigger
Class: DNA
Position: chr7:74650211-74650770

Validation of *GTF2IRD2B* isoform by an orthogonal data type (Sanger)



UCSC Genome Browser snapshot for hg38 chr7:75,090,834-75,150,884 displaying (from top):

Black: *GTF2IRD2B* isoform described in this study

Black: "Your sequence from BLAT results" is Sanger sequence generated from the validation PCR product

Red/Orange: Oxford Nanopore (ONT) direct RNA-seq reads

Blue: RefSeq and UCSC Genes

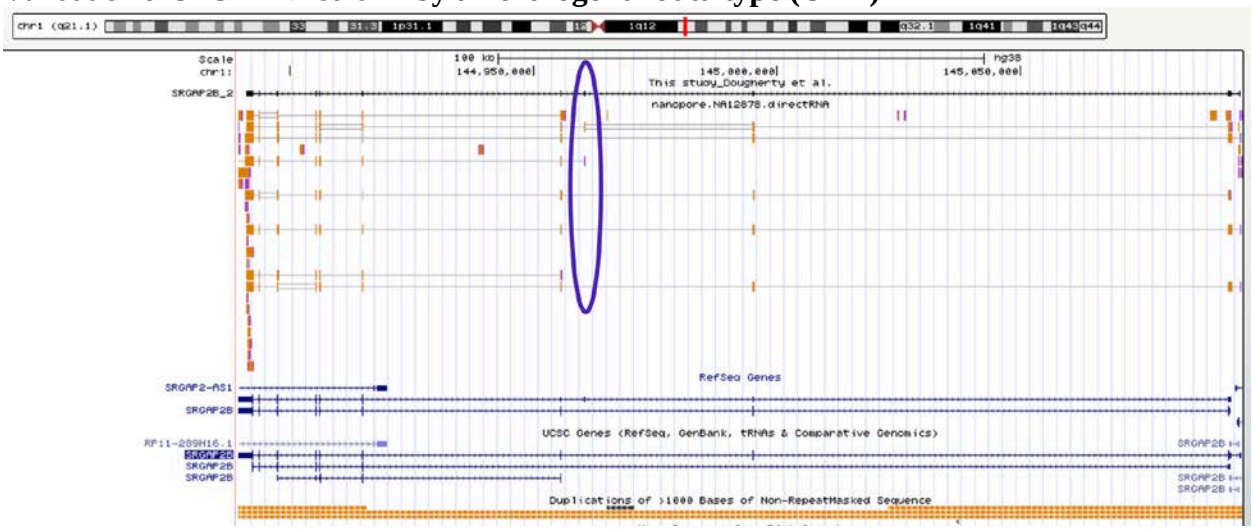
Orange: Segmental duplication blocks (UCSC color code for 98-99% sequence identity)

Black: Repeat elements by RepeatMasker

Summary: **This study uncovered a new isoform from the gene family *GTF2IRD2*, with a new first exon encoded by a repeat element of the TcMar-Tigger7 DNA family.** To validate this first exon, we generated oligonucleotide primers against the Tigger exon (5') and in exon 4 of *GTFIRD2* (3'). The PCR product amplified from brain cDNA was Sanger sequenced from the 3' primer and the BLAT results are shown. The sequence is definitely derived from *GTF2IRD* (and not the ancestral *GTF2I*), we cannot assign whether it is derived from the *GTF2IRD2* or *GTF2IRD2B* locus due to a lack of paralog-specific nucleotides. The ONT data are abundant in this region and a subset of the reads (~50%) are shown. No ONT reads were found that support the Tigger7 exon splicing to exon 2 of the canonical *GTF2IRD2* isoform shown currently in RefSeq. In the RepeatMasker track, the repeat element at the coordinates of the first exon is:

Name: Tigger7
Family: TcMar-Tigger
Class: DNA
Position: chr7:75094267-75095000

Validation of *SRGAP2B* isoform by an orthogonal data type (ONT)



UCSC Genome Browser snapshot for hg38 chr1:144,889,249-145,095,824 displaying (from top):

Black: *SRGAP2B* isoform described in this study

Orange: Oxford Nanopore (ONT) direct RNA-seq reads

Purple oval represents the 61 nucleotide exon described in lines 336-349 and depicted in Figure 6 of the manuscript.

Blue: RefSeq and UCSC Genes

Orange: Segmental duplication blocks (UCSC color code for 98-99% sequence identity)

Summary: **This study uncovered a new isoform of *SRGAP2B*, with a different first exon than RefSeq annotation and containing a 61 nucleotide (nt) exon that results in ORF truncation.** The ONT data are sparse in this region, perhaps due to low expression levels in this cell line. However, among the data here, two ONT reads (one ~full-length read and one partial) support the inclusion of the 61nt exon. The sequence identity level is likely too high in this duplication block to confidently assign these reads to one *SRGAP2* paralog, but our data supports assignment of these two reads specifically to the *SRGAP2B* copy since inclusion of this exon is only seen among this paralog's transcripts. (Figure 6 explains the paralog-specific splice site change as rationale.) Note that the new isoform combines features of two known isoforms derived from this paralog. The RefSeq track shows an isoform that includes the 61nt exon, but it does not harbor the same promoter as our new isoform. Conversely, the UCSC Genes track does display an isoform with the same promoter as our new isoform, but it does not include the 61nt exon.