scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Integrating Iso-seq and RNA-seq data for the reannotation of the killifish telencephalon transcriptome

Rajagopal Ayana[1,2] ✉, Tatiana Krutikhina[1], Jolien Van houcke[1], Caroline Zandecki[1,2], Valerie Mariën[1], Gregory E. Maes[3,4], Eve Seuntjens[2,5] ✉ & Lutgarde Arckens[1,5] ✉

The short-lived and rapidly aging African turquoise killifish, *Nothobranchius furzeri* GRZ is a unique model to study vertebrate aging. Current genomic and full-length transcriptomic sequencing lacks full gene annotations, resulting in poor mapping in bulk and single-cell transcriptomic studies. In our efforts to reannotate the transcriptome of the killifish telencephalon, we combined long-read (Smrt-Isoseq) and short-read transcriptome sequencing approaches. A total of 17,008 full-length isoforms, including 6763 novel ones were obtained (51 bp to 7,500 bp). The killifish telencephalon comprises 25% multi-exon genes, while over 50% are mono-exon genes. We discovered novel non-coding and coding sequences in both young and aged telencephali. We integrated long-read and RNA-seq data to construct a comprehensive transcriptome and profiled expression dynamics across the aging telencephalon. Our gene models demonstrate greater detail and accuracy than Ensembl, with more precise polyA locations. Alternative splicing analysis revealed 29 events altered with aging, which involved changes in ribosome function, gap junction and mRNA surveillance pathways. These generated resources pave the way for future functional genomic studies in this biogerontology model.

Background & Summary

High-throughput transcriptomics enables studying overall aging-related changes, to reveal the molecular changes in cellular pathways, thus preparing a base for functional studies of specific proteins and pathways, and, eventually, pharmacological interventions[1,2]. Most studies are fundamentally reliant on accurate and complete gene annotations, which are only possible based on high-quality reference genomes. In species with incomplete or poorly annotated genomes such as many teleost fish, short-read (SR) and long-read (LR) sequencing technologies must often be integrated to resolve transcript diversity and capture complex splicing patterns. On the one hand, SR RNA-sequencing (RNA-seq) offers high depth and base-level accuracy for identifying expression changes in both coding and non-coding RNA. On the other hand, third-generation LR DNA sequencing technologies, such as Oxford Nanopore Technologies and Pacific Biosciences' SMRT Sequencing enable full-length transcript identification and novel isoform discovery. These complementary approaches are especially critical in fish species[3], where genome complexity includes high levels of gene duplication and repetitive regions thus posing hurdles for accurate transcriptome assembly and annotation. For example, despite initial transcriptomic resources generated from Illumina and Roche 454 platforms using the medaka genome as a scaffold, the African turquoise killifish (*Nothobranchius furzeri*) still lacks a complete and tissue-specific transcript annotation, particularly for the central nervous system[4].

Nothobranchius furzeri (N. furzeri GRZ), the African turquoise killifish, is the shortest living vertebrate in captivity and serves as a well-characterized animal model. Its remarkable and age-dependent regenerative

[1]Laboratory of Neuroplasticity and Neuroproteomics, Animal Physiology and Neurobiology Section, Department of Biology, KU Leuven, Leuven, Belgium. [2]Laboratory of Developmental Neurobiology, Animal Physiology and Neurobiology Section, Department of Biology, KU Leuven, Leuven, Belgium. [3]Centre for Human Genetics, Genomics Core, UZ-KU Leuven, Leuven, Belgium. [4]BioGenomics-Consultancy, Diestesteenweg 185, Linden, Belgium. [5]KULeuven, Leuven Brain Institute, Leuven, Belgium. ✉e-mail: rajagopal.ayana@kuleuven.be; eve.seuntjens@kuleuven.be; lut.arckens@kuleuven.be

abilities make the killifish a competitive model to explore new neuro-reparative strategies that can be translated to the mammalian central nervous system (CNS)[5]. With respect to the molecular, cellular, and integrative traits of aging, *N. furzeri* shows significant similarities to mammals, including humans[6–8]. In addition, a rapid decline in neurogenesis and increased neurodegeneration akin to humans is observed[6,8]. As our understanding of killifish transcriptomic complexity improves, it is apparent that existing gene annotation, principally originating from four sources (RefSeq, GENCODE, NFINgb, Ensembl), remains incomplete. In the brain, and especially the telencephalon, *N. furzeri* exhibits a pronounced age-dependent decline in neurogenesis and increased neuronal loss. However, the currently available transcriptome annotations for killifish remain incomplete, particularly for brain-specific genes and isoforms relevant to neurogenesis and neurodegeneration. These limitations hinder mechanistic insights into neural aging and restrict the utility of *N. furzeri* for modelling neurodegenerative diseases. Notably, as seen in human transcriptomics studies, incomplete annotations disproportionately affect genes with complex splicing or brain-enriched expression patterns[9]. Here, we aimed to increase the resolution of the current transcriptome annotations for the killifish telencephalon to fuel neuroscience research in the killifish and the many downstream health applications.

We addressed this gap by improving transcriptome annotations of the killifish telencephalon using an integrative short- and long-read RNA sequencing approach. By focusing on this brain region, homologous to the mammalian forebrain and central to executive functions, learning, memory and behavioural flexibility, we aim to enable more accurate transcript-level resolution of gene expression and splicing dynamics, thereby enhancing the utility of the killifish as a translational model in neuroscience and aging research. In this study, we built a high-quality custom reference transcriptome for the female killifish telencephalon (GRZ strain) using PACBIO SMRT-seq and Isoseq-based analyses. First, we analysed age-related differential gene expression in young vs. old killifish telencephali at the bulk level and elaborated on the molecular pathways that were enriched during aging. Combining this expression information with the SMRT-seq data, we modelled the transcriptomic and isoform dysregulation upon aging. This way, we constructed the complete transcriptome specific to the telencephalon (combined young/aged), and classified the various isoforms identified. Finally, we tested for alternative splicing events occurring in differentially expressed genes upon aging. The obtained long-read dataset is ideal for improving the existing killifish gene annotation file for future mapping single-cell and bulk RNA-sequencing datasets.

Methods

Killifish breeding. All experiments have been performed on the fast-aging *N. furzeri* of the inbred GRZ strain, with a mean lifespan of 4–6 months[5]. This strain shows fast juvenile growth and very early reproduction, followed by an accelerated aging process (compared to longer-lived strains, such as the MZM strain). Given the sex-dependent variation in growth and aging rates, all experiments were performed exclusively on young-adult 7-week- and aged 17-week-old *N. furzeri* female fish. Three females and one male were group-housed in 3.5 L tanks, containing a non-transparent T-tube for cage enrichment, in a multilinking Tecniplast ZebTec aquarium system under standardized conditions: temperature 28 °C, pH 7, conductivity 600 µs, 12 h/12 h light/dark cycle. From hatching, the fish are fed twice a day with *Artemia salina* (Ocean Nutrition), 2 weeks after hatching, mosquito larvae (*Chironomidae*) were added to their diet. From 5 weeks onwards, the fish were only fed with mosquito larvae. All experiments were approved by the KU Leuven ethical committee in accordance with the European Communities Council Directive of 20 October 2010 (2010/63/EU).

SMRT sequencing library preparation. According to the lifespan experiments in the lab, 7-week old female fish are classified as sexually mature young adults (100% survival in colony), while the 17-week fish are considered aged (76% survival in colony; comparable to a human age of ~75 years)[5]. Per age, telencephali of three fish were pooled after dissection and snap freezing on dry ice. The whole RNA was extracted using the RNAEasy Micro kit (Qiagen). Final RNA quality and integrity were assessed using the DNA 12000 kit on Bioanalyzer (Agilent). The SMRT sequencing was performed on the PACBIO Sequel at the Genomics Core at KU Leuven (Belgium) using the Clontech SMARTer PCR cDNA Synthesis Kit (recommended by PACBIO). Multiple parallel PCR reactions were set after the first strand synthesis employing the PrimeSTAR GXL DNA Polymerase kit. Next, cDNA of a specific length was selected by splitting these PCR reaction products into two fractions, purifying one with 1X AMPure PB Beads and the second with 0.4X AMPure PB Beads, and subsequently mixing both purified products in an equimolar fashion. These libraries were constructed with 20 h movie-time using the SMRTbell template prep kit 1.0., and sequenced employing Sequel Sequencing, and Binding kits 3.0 on a Sequel I platform with SMRTCells V3.0 LR[10]. The mRNAs were selected by poly-A tails, without 5′cap selection, which was considered in the subsequent analysis. The reference genome, gene annotation, and GO terms were downloaded from the *N. furzeri* Information Network Genome Browser[11] (NFINgb).

PACBIO Isoseq3 pipeline. Long-read platforms have some limitations. They are low-throughput unlike short-read sequencing methods, resulting in reduced coverage for low-abundance or short transcripts (e.g., some lncRNAs or short UTR isoforms). The method also poses a 3′ bias in cDNA capture, which may affect complete TSS representation. There could also be higher per-base error rates, particularly indels and mismatches in homopolymer regions. To address these issues, we used the Isoseq3 clustering pipeline followed by quality filtering and applied SQANTI2 to evaluate transcript integrity and junction support, using short-read Illumina data as a reference.

The analysis was performed with the SMRT Link 8.0 IsoSeq3 analysis pipeline. The full Isoseq3 pipeline and the parameters for the tools were chosen according to the Isoseq3 user manual. The first step of this pipeline assembles circular consensus reads (CCS) from raw subreads with the *ccs* command. We used *ccs* version 4.0.0 with options "--noPolish --minPasses 1". Next, the primer sequences were trimmed/clipped, and

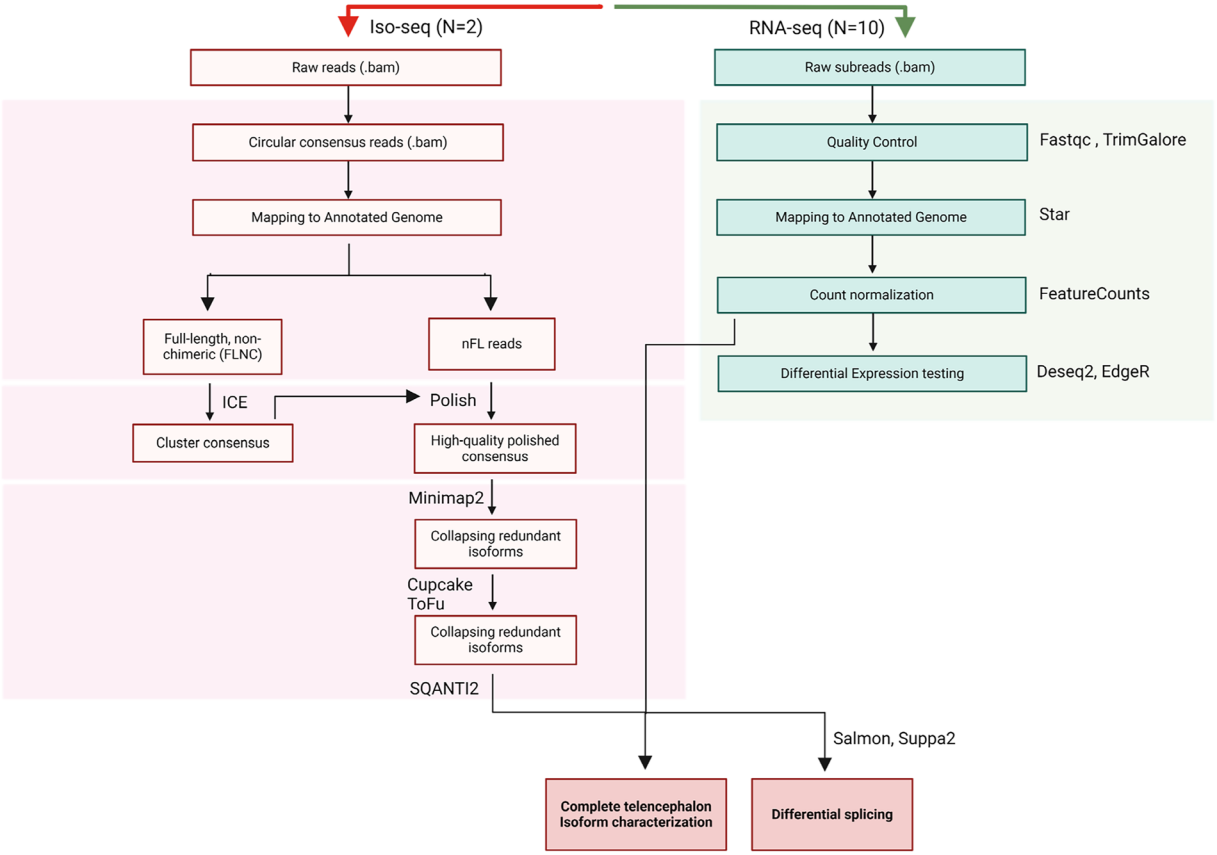


Fig. 1 The pipeline to combine differential expression analysis and ISO-seq data.

reads were demultiplexed using the *lima-isoseq* command under default settings. *Lima-isoseq* command was employed for removing unwanted combinations and orienting sequences in 5′-3′ direction, creating full-length non-concatemer (FLNC) reads. The reads were classified as full length if they had the 5′ and 3′ primers and the poly-A tail. In our samples, 84% (7-week) and 80% (17-week) of CCS passed through *lima* filters, which confirms a good quality of the library preparation. The transcripts were refined, clustered, and polished with the Isoseq3 v. 3.2.2. The following commands and parameters were applied to process the data: (1) *isoseq3 refine* with options "--require-polya"; (2) i*soseq3 cluster* with default settings; (3) *isoseq3 polish* with default settings. The high-quality isoform sequences were obtained by processing the data through the described pipeline (Fig. 1). The number of transcripts drastically decreased after the clustering step. The mean length of polished transcripts equaled 2,977 bp in the 7-week sample and 3,427 bp in the 17-week sample. Statistics regarding number and distribution of gene isoform differences between 7- and 17-week samples are catalogued (Suppl. Figure 1a–c). The length distribution of transcripts in 7-week vs 17-week samples passing through the Isoseq pipeline was recorded (Suppl. Figure 1d,e).

Post isoseq analyses. To ensure the generation of a high-confidence transcriptome suitable for downstream analysis, we employed a multi-step filtering pipeline leveraging both Cupcake ToFu followed by SQANTI2, which target typical errors in long-read transcript assemblies such as redundancy, degradation artifacts, and structural inaccuracies.

Mapping. The reads of the *N. furzeri* transcriptome[4,11] were mapped with *Minimap2* v. 2.17[12] with settings "-ax splice -t 30 -uf --secondary = no -C5", which are compatible with subsequent Cupcake Tofu analysis[13]. This mapper supports both spliced and unspliced alignment to the reference genome. Here, the parameter "-ax splice" means that the input reads are spliced, and "--secondary = no" means that mapping is allowed only to one location in the genome. 99% and 99.3% of reads were uniquely mapped to the genome in 7- and 17-week samples, respectively.

Collapsing redundant isoforms, filtering 5′-degraded ones. Next, *Cupcake ToFu* was employed for collapsing redundant, or representing the same sequence, isoforms. The alignments with 99% coverage and 95% identity were selected, and the isoforms were collapsed by the *collapse_isoforms_by_sam.py* script with settings "--dun-merge-5-shorter", meaning that slightly different on the 5′ end isoforms are collapsed into the longer ones. Since the library preparation did not include 5′-cap selection, the 5′-degraded transcripts can be a technical artifact and have no biological function. So, we collapsed 5′-degraded isoforms into longer ones with the

filter_away_subset.py and quantified with the *get_abundance_post_collapse.py* script. These scripts are designed to retain only the most complete isoform at each locus, reducing noise caused by incomplete cDNA synthesis or RNA degradation. After collapsing isoforms, we retained 30,943 isoforms at 7-week and 24,706 isoforms at 17-weeks, which were further reduced to 17,274 and 15,919 isoforms, respectively, after filtering out transcripts with incomplete 5′ ends. With the *chain_samples.py* (again, with "--dun-merge-5-shorter" parameter), two isoform datasets were chained into one .gtf file.

Annotation and structural characterization of the isoforms. We annotated the isoforms with *SQANTI2* v. 7.4.0[14] using the script *sqanti_qc.py*. It returns the reference-corrected transcriptome, classification file, and a set of splice junction descriptors. The open reading frame (ORF) predictor, embedded in this tool, differentiated coding from noncoding reads. The nonsense-mediated decay (NMD) targets were also predicted. Next, we filtered out the isoforms, which are likely to be library preparation artifacts, with the *sqanti_filter.py* script. It uses the transcript-level descriptors and the reference-corrected transcriptome and filters out transcripts with signs of reverse transcriptase template switching and intra-priming. *Sqanti_filter.py* retains sequence according to the criteria: a) Its junctions fully match the reference genome (full splice match), and the 3′ end is reliable (the reliability of the 3′ end is calculated on the previous step). b) It is not a full splice match, but three conditions are met. (1) no reverse transcriptase switching signs are noticed; (2) the 3′ end is reliable; (3) all splice junctions are canonical or are confirmed by at least three short read transcripts (coverage threshold). The splice sites are canonical if they have the structure GC/AG or AT/AC. All the other dinucleotide combinations are considered non-canonical.

Further using the *sqanti_filter.py* script, we filtered out transcripts that met any of the following conditions: a) splice junctions unsupported by short-read seq evidence, b) contains non-canonical splice junctions without strong expression, c) flagged as potential internal priming events, d) showed strong RT-switching signatures, e) were mono-exonic, low-expression, and antisense or intergenic (often artifacts). We retained: a) FSMs and NIC/NNC isoforms with valid canonical junctions and/or strong short-read support, b) Multi-exonic transcripts with plausible ORFs or transcript models supported by annotation. This step removed a significant proportion of low-quality isoforms, improving transcriptome reliability without losing real biological diversity. There is no other published data available that provides more guidance for the validation of these variants. Therefore, we took several genes, for which the number of isoforms is known (for example, the CNR1 gene has three known isoforms in mammals), and checked the identified number of isoforms for them. The non-filtered version of our data demonstrated nine isoforms for CNR1, which differed in the number of exons. The filtered version included four isoforms, of which one is known, and the other three are novel. These results confirmed that the 5′-filtered variant of .gtf is more reliable. The .fasta and .gtf files of chained filtered transcriptome were used in subsequent analyses. The complete pipeline combining long-read and bulk RNA-seq data has been summarized (Fig. 1).

Integration optimization strategy. To minimize the impact of long-read sequencing errors on transcript integrity, we used above optimization steps: 1. *IsoSeq3 clustering and polishing*: Raw subreads were clustered into full-length non-concatemer (FLNC) transcripts and polished to consensus accuracy. 2. *Redundancy removal*: Cupcake ToFu was used to collapse isoforms and filter 5′ degraded variants. 3. *SQANTI2 filtering*: Transcripts were structurally classified and filtered using short-read-supported splice junction validation. 4. *Expression cross-validation*: Only isoforms with consistent short-read expression and/or supported junctions were retained for downstream analysis. Together these steps ensure that the final chained transcriptome is comprehensive and biologically reliable, particularly in the context of analyzing age-related changes in neurogenesis and brain transcriptome complexity.

Alternative splicing analysis. We mapped and counted the quality-controlled short reads on 7- and 17-week samples to the chained transcriptome with *Salmon* tool[15] using parameters "--seqBias --gcBias -l IU" (assuming that the library is unstranded and performing the sequence-specific and GC bias correction). This gave us the expression of transcripts in transcripts per million (TPM). Next, we performed the differential splicing analysis using the *SUPPA* program to assess the Proportion or Percentage Spliced in (PSI) values[16]. The *Suppa.py generateEvents* created the file with the differential splicing events, based on the chain transcriptome file (.gtf). Next, with the *Suppa.py psiPerEvent*, we counted the event inclusion level for each splicing event, and with the *Suppa.py diffSplice*, we statistically compared the abundance of splicing events per isoform between the age groups. The resulting Dpsi file gives information about the deltaPSI which is the event PSI difference between old and young (ΔPSI = PSI_young - PSI_old) and the p-value of this deltaPSI, while psivec gave the PSI-value for each event per condition. We further used the cut-off of p-value ≤ 0.05 and recommended cutoff for biological significance: $|\Delta PSI| \geq 0.1$ (i.e., a 10% change in splicing inclusion/exclusion). Further we also examined event inclusivity by filtering based on transcript expression across replicates (TPM > 1). These robust thresholds are generally considered biologically meaningful, particularly for exon skipping or intron retention events. We obtained 29 differentially spliced genes, for which an MA plot was created (Fig. 5a). Overall, this ensures that PSI values are not derived from low-confidence transcript quantifications.

Pathway analyses of the 29 differential genes were performed using ShinyGO[17] which is mapping efficient for non-model species such as the killifish. We considered robust thresholds of adjusted p-value (FDR) < 0.05, including KEGG pathways only and minimum number of genes enriched in a pathway $= 1$.

Identifying differentially expressed genes using short-read sequencing data. To validate long read data and expressed transcripts, we performed Quant-seq (RNA-seq) of young and aged killifish telencephali. Telencephali from five female fish of 7- and 17-weeks of age were collected. We extracted RNA using the RNAEasy Micro kit (Qiagen). Total RNA-sequencing of 10 telencephalic samples (5 per condition) was

Data Type	Tissue	Biosample Accession	SRA Accession
PACBIO IsoSeq (Transcriptomic Data)[19]	Young	SAMN45173805	SRR31611110
	Aged	SAMN45173806	SRR31611109
Illumina Short reads (RNASeq)[19]	Young (1–5)	SAMN45165978 SAMN45165979 SAMN45165980 SAMN45165981 SAMN45165982	SRR31602945 SRR31602944 SRR31602943 SRR31602942 SRR31602941
	Aged (1–5)	SAMN45165983 SAMN45165984 SAMN45165985 SAMN45165986 SAMN45165987	SRR31602940 SRR31602939 SRR31602938 SRR31602937 SRR31602936
Isoseq Assembly	Figshare repository	[20]	
Isoseq Predicted genes	Figshare repository	[20]	

Table 1. List of data generated in the current study and their corresponding accession numbers.

performed by Quant-seq (Lexogen$^{TM/R)}$ and further analysis was done. Sequence libraries were prepared with the Lexogen QuantSeq 3′ mRNA-Seq library prep kit according to the manufacturer protocol. Samples were indexed to allow for multiplexing. Library quality and size range was assessed using a Bioanalyzer (Agilent Technologies) with the DNA 1000 kit (Agilent Technologies, California, USA). Libraries were subsequently sequenced on an Illumina HiSeq 4000 instrument. Single-end reads of 50 bp length were produced with a minimum of 1 M reads per sample. The quality of sequenced reads was checked using *FastQC* tool on the Galaxy bioinformatics platform, all base calls with Phred score below 20 were removed. Further, the quality of the reads was improved by removing the Illumina standard adapters, the reads with length below 50 bp, and poly-A tails. A trimming of 12 base pairs at the 3′ end was performed by the parameter "--three_prime_clip_R1 12" and followed by a new quality check integrated in this command with the option "--fastqc". We used the chained Isoseq transcriptome, as used in previous analyses, to map the transcripts to ensure highest mapping scores. Using *STAR*[18], we mapped the raw data to the specific transcriptome which resulted in unique mapping percentages between 48.8–67.2%. These mapped .bam files and the chained transcriptome were used for raw counting of the mapped reads by using *Featurecounts* (default settings). The output files comprise of a table which includes the geneID, chromosome number, gene counts and the effective gene length. To visualize any difference in expression between the young (7-week) and the old (17-week) data, a differential expression analysis was subsequently performed. We performed data normalization, log fold change estimation, statistical testing, filtering along with employing multiple testing correction methods using *edgeR* v. 3.24.

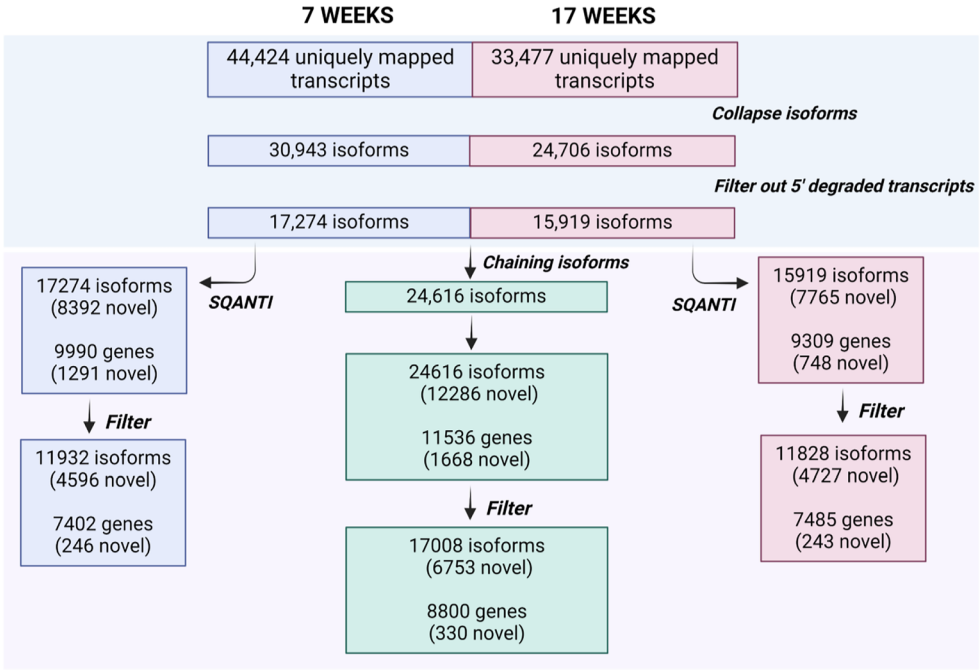
Data Records

All RNA-seq raw reads of killifish telencephalon were deposited on NCBI SRA under project accession number: PRJNA1194046/SRP549330[19]. The PACBIO Isoseq raw reads files are also deposited there (Table 1). The complete Isoseq assembly transcriptome and the predicted gene sequence files (main dataset) are available on Figshare[20]. In detail, this repository houses the PACBIO Isoseq assembly and predicted gene sequence and annotation files of the female *N. furzeri* telencephalon.

Technical Validation

PACBIO Long-read sequencing identifies several thousand isoforms. We extracted total RNA from killifish telencephalon that shares homologies to all four mammalian pallial areas. Using the Iso-Seq3 pipeline and SQANTI filtering, we recovered a final Iso-Seq transcriptome composed of 30,943 and 24,706 uniquely mapped gene-isoforms in young and aged killifish telencephali respectively (Fig. 2a). Post 5′ degraded transcripts filtration, we ended up with 17,274 and 15,919 isoforms in young and aged fish, respectively. These 17,274 high-quality isoforms cover 9,990 genes in the 7-week killifish telencephalon, and 15,919 high-quality isoforms in the 17-week killifish telencephalon, covering 9,309 genes. The quality control parameters and the splice junction percentages were very similar in 7- and 17-week samples (Suppl. Figure 2a–c). We used SQANTI2 to characterize this transcript diversity in the telencephalon and improve the cell type characterization in our single-cell study[10]. At the end of the SQANTI2 pipeline, three telencephalon-specific classification files were obtained (1) 7-week sample, (2) 17-week sample, and (3) their chained version. We chained the two transcriptomes to make a versatile catalogue for the killifish telencephalon. This resulted in a combined gene-isoform diversity of 24,616 gene-isoforms that includes 12,286 novel isoforms, 11,536 genes and 1668 novel genes in total. This indicates that 72.8% of isoforms are shared between the two ages. In comparison, the Ensembl full transcriptome of killifish (release 107, Assembly Nfu_20140520) contains 23,428 genes and 57,281 isoforms. With the Iso-seq transcriptome, we identified a similar number of genes, but disproportionate number of isoforms, suggesting that alternative splicing and/or alternative transcription start/termination sites (TSSs and TTSs) may be less pervasive in the telencephalon than predicted through the Ensembl annotations. We identified 12,286 novel isoforms that did not match previously annotated Ensembl isoforms expanding the resolution of the transcriptome immensely. Furthermore, within known annotated genes we observed a higher number of alternative isoforms (Suppl. Figure 2b). There was a higher percentage of genes that had two or more isoforms in the Iso-Seq transcriptome compared to the Ensembl transcriptome (Ensembl: 24%; Iso-Seq: 35%).

a)



b)

Age	Raw subreads	CCS	FLNC	Unpolished clusters	Polished transcript candidates	High-quality isoforms
7 weeks	15,180,383	461,340	379,986	44,853	44,841	44,695
17 weeks	12,324,737	409,487	329,312	33,837	33,827	33,706

Fig. 2 (**a**) The number of transcripts on different stages of the Isoseq3 pipeline. (**b**) Detailed description of the pipeline and filtration steps involved before chaining the transcriptomes.

These results demonstrate how filtering substantially reduced transcriptome complexity, removing 30–45% of total isoforms, while retaining high-confidence, full-length transcripts essential for robust downstream analyses. Notably, filtering prevented the overestimation of isoform diversity, which would have otherwise confounded conclusions about transcriptomic remodelling in aging and neural function. By adhering to the best practices provided by the Cupcake ToFu and SQANTI2 frameworks, and validating transcript structures with short-read support, we produced a well-curated and biologically interpretable isoform catalogue for the turquoise killifish telencephalon across age[20].

Iso-seq3 pipeline constructs a state-of-the-art telencephalon specific transcriptome. The whole tissue was dissociated and both cytoplasmic and nuclear RNA was collected for long-read RNA sequencing (SMRT-seq/Iso-seq). On average, we found 435,413 (7-week: 461,340, 17-week: 409,487) circular consensus sequences (CCS) to be produced from the raw subreads between the two fish ages (Fig. 2b). These CCS were then filtered, clustered and polished into full-length transcripts producing an average of 39,334 (7-week: 44,841, 17-week: 33,827). The summary of the gradual changes in the number of transcripts and distributions of their length is presented (Fig. 2b, Suppl. Figure 1a). The number of transcripts drastically decreased after the clustering step. The mean length of polished transcripts was equal to 2,977 base pairs in 7-week sample and 3,427 base pairs in 17-week sample. We then tested the quality of the transcriptome using the aligner minimap2, due to its compatibility to subsequent Cupcake Tofu analysis. Minimap2 identified 44,695 and 33,706 high quality isoforms in young (7-week) and aged (17-week) killifish respectively (Fig. 2b).

Classification of isoforms based on long-read analysis. Within the complete dataset, the largest proportion of the transcripts were known protein-coding genes (Fig. 3a). The least common type of transcript is intergenic (<1%), and the most common are full splice match (61%) and incomplete splice match (18%), which together are classified as known transcripts (Fig. 3a,b). At the same time, novel transcripts demonstrated all possible variants of splice junctions, including intergenic, antisense, genic genomic, novel in a catalog, novel not in catalog (Fig. 3c). Novel isoforms represented nearly 40% of all transcripts, with 87% of them expressed from known genes (Fig. 3c). There is minimal difference between 7- and 17-weeks samples in the percentage of novel isoforms

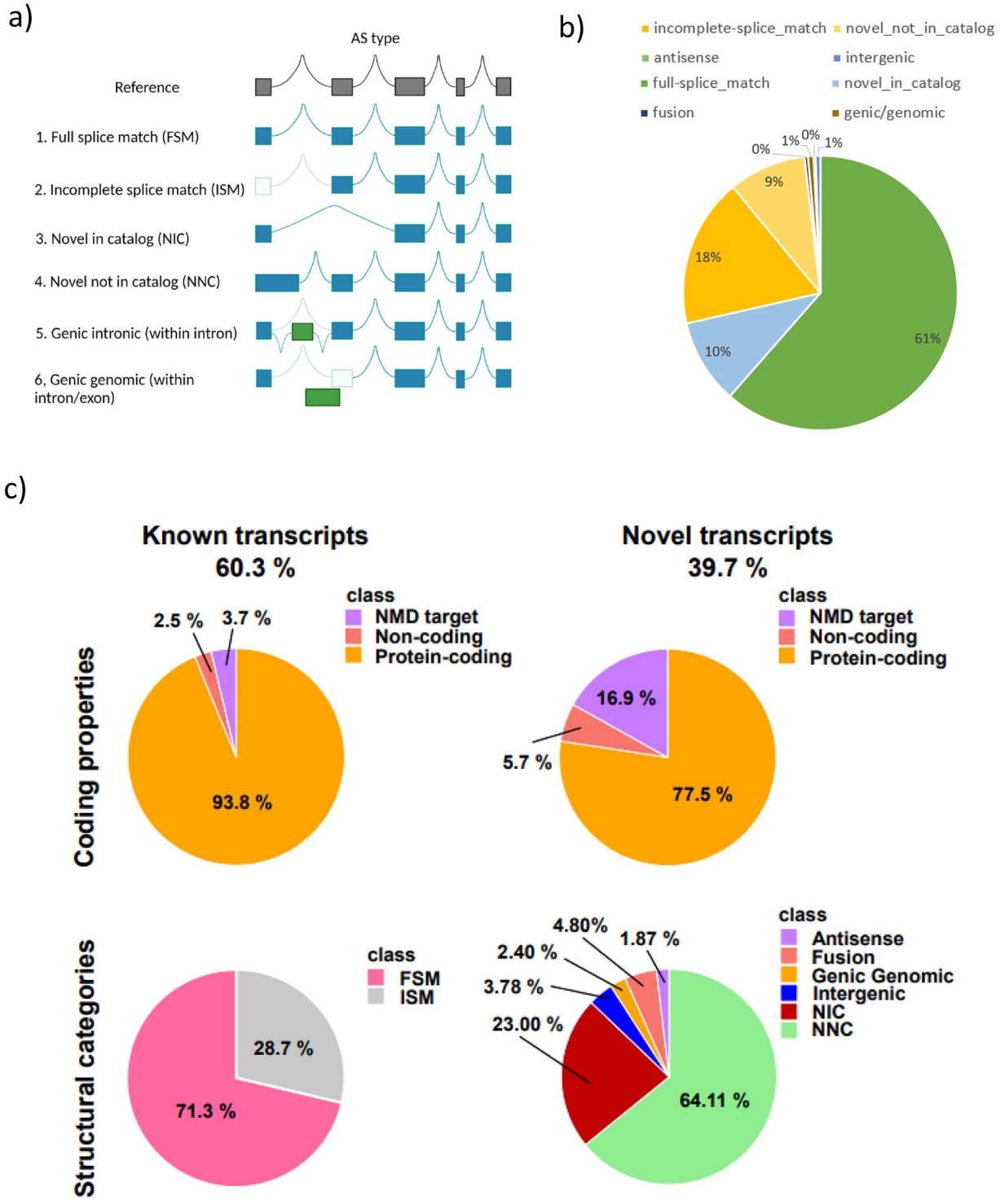
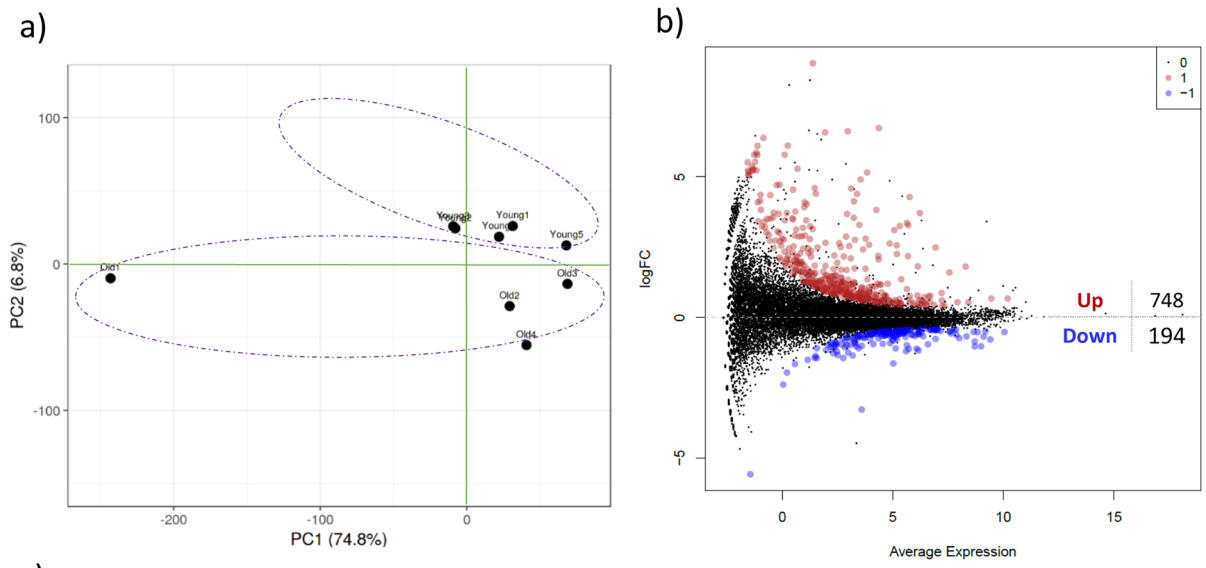
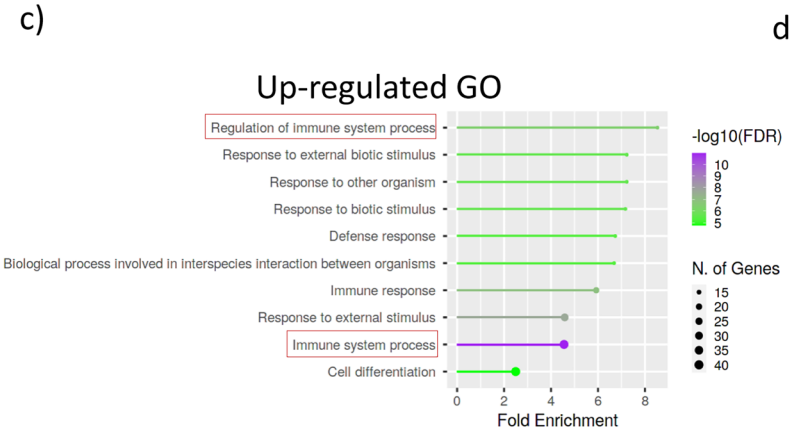


Fig. 3 (**a**) The reference catalog of alternative spliced (AS) types existing in the telencephalon Isoseq transcriptome. (**b**) Pie chart showing the distribution of the different AS types and their prevalence in the final transcriptome. (**c**) Detailed classification of known and novel isoforms retained from the final transcriptome using SQANTI2.

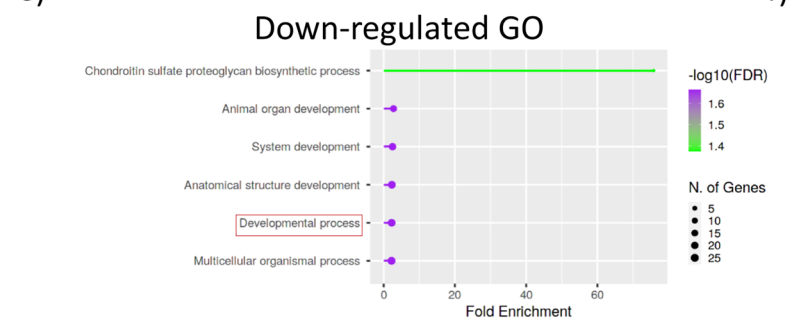
(38.5% and 39.9% respectively). The novel and known transcripts were summarized by coding properties and splice junctions (Fig. 3c). Among the novel genes and isoforms, about 17% is non-coding, and 5.7% of coding mRNAs are labeled as NMD targets, meaning that they contain premature stop codon and are degraded by the nonsense-mediated decay machinery. This is an increase when compared to ratios in the known transcripts. The NMD targets demonstrated less FLNC supporting reads than non-NMD mRNAs, according to the Student T-test ($p = 0.026$ in the 7-week sample, $p = 0.039$ in the 17-week sample). The aberrant alternative splicing may lead to





d)

IFN pathway related genes		
Gene name	log2FC	P-value
IRF3	2.22	2.81E-04
CD68	2.76	3.07E-03
AIF1	1.01	1.33E-03
PFN1	1.75	5.20E-07
B2M (1OF2)	2.85	3.41E-03
CCL-C5A	6.49	8.78-04
DHX58	1.95	6.105-04
IFIH1	3.33	1.31-04
PSMB10	3.31	4.99E-04
RSAD2	4.37	4.58E-04
TLR3	1.62	7.82-03
TNFRSF21	0.59	2.318-02
TRIM25	3.08	2.15E -03
ZNFX1	3.28	3.02E -03



f)

Proliferation related genes		
Gene name	Log2FC	P-value
KIAA0101	-1.74	1.25E-04
MARCKS	-0.51	1.59E-02
HES2	-4.06	7.66E-03
NFIXB	-3.95	1.65E-02
MSI	-1.77	3.70E-03
MCM4	-1.16	7.34E-03
MK167	-0.93	1.13E-02
GBX1	-0.87	2.82E-02

Fig. 4 (**a**) PCA plot showing order of 10 samples from bulk RNA-seq experiment. (**b**) Volcano plot showing up and down-regulated genes upon aging; FD cut off: ±1. (**c**) GO terms associated with Up-regulated gene list (>1.2). (**d**) Associated list of genes, associated with IFN and immune modulation. (**e**) GO terms associated with down-regulated gene list (<−1.2). (**f**) List of genes associated with proliferation.

NMD targets[21], and the percentage of NMD targets does not differ between young and old killifish samples (8.4 and 7.7% of transcripts respectively). The non-coding transcripts also include the long non-coding RNAs (lncRNAs), which have a length ranging from 182 to 1776 bp with no open reading frame (ORF), and thus, no translation into proteins. Most of the lncRNAs have a poly-A tail, so they are not filtered out during library preparation. Another transcript category is "gene fusion", which are chimeric transcripts, named for example, "*map2_kansl1l*". They are expressed from fusion genes or are encoded by two different genes and joined by trans-splicing. Most

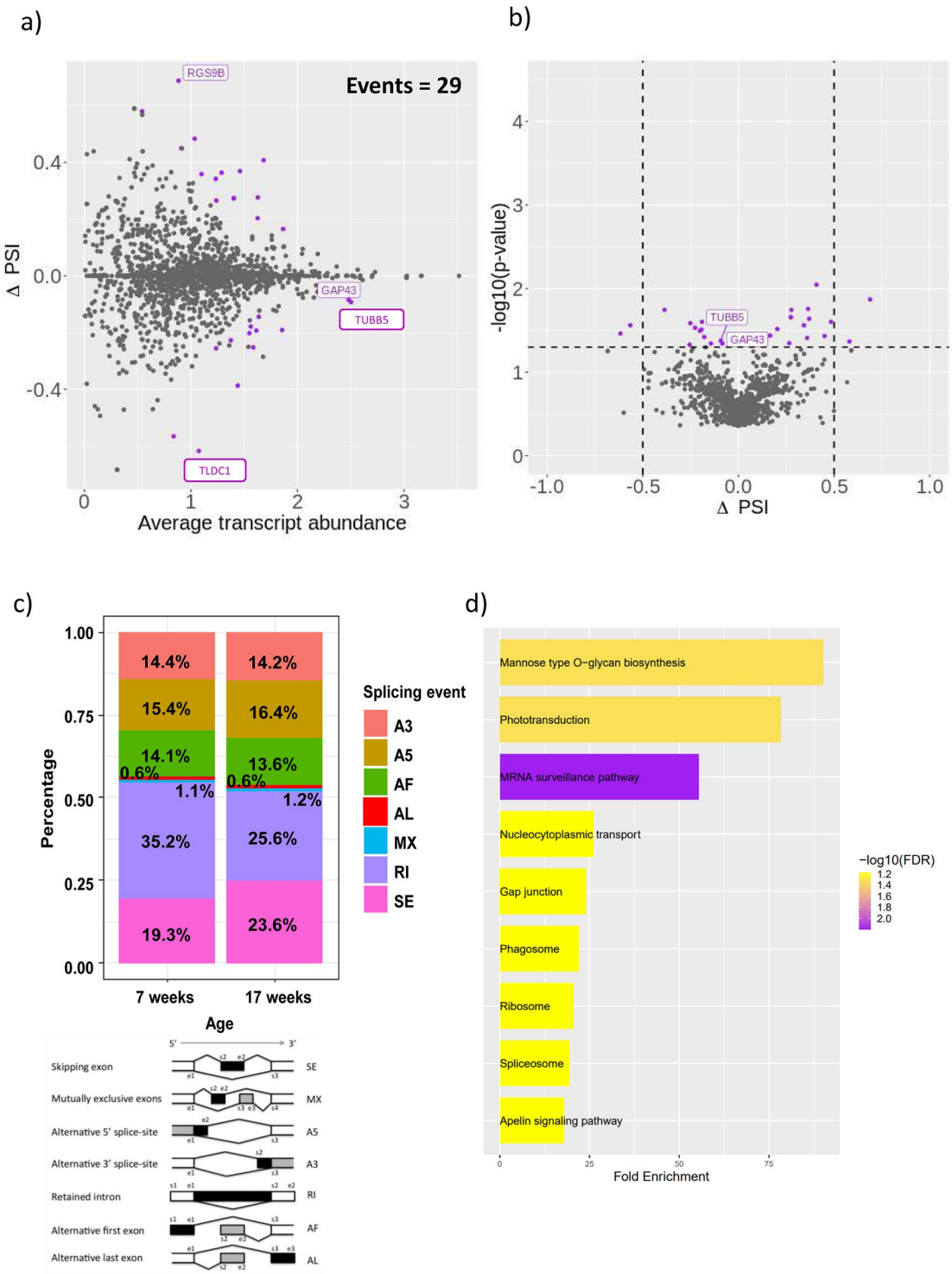


Fig. 5 (**a**) Suppa2 derived alternative splicing events across aging. In this graph, we ordered average transcript abundance with PSI values. Note: Purple dots indicate significance. (**b**) In this graph, we ordered PSI values with log(P-values). Note: Purple dots indicate significance. (**c**) Fraction and type of AS transcripts that undergo different types of splicing events, based on the long-read data. Note: Nomenclature for alternative splicing events in Suppa2; For the regions filled with black color, the relative inclusion level of the splicing event (PSI value) is calculated. The grey region represents the other, alternative variant. (**d**) Dysregulated pathways among the list of 29 genes that were alternatively spliced. Note: Significance order: Purple to Yellow.

of them belong to the multi-exon subcategory containing, on average, about 10.9 exons per transcript. Some additionally have retained introns. Seventy three percent of the "gene fusion" transcripts have an ORF and are not NMD targets, so, they are coding for a chimeric protein. Overall, we built and updated the diversity of transcripts specific to the telencephalon.

Differential expression changes upon aging. To establish the differential splicing usage upon aging, we first profiled the telencephalic transcript expression using bulk RNA-seq to characterize the aging-induced molecular alterations using the young adult and aged killifish. Principal component analysis (PCA) revealed a gradual transcriptomic shift between the two conditions (Fig. 4a). We identified 748 genes upregulated and 194 downregulated upon aging (cut-offs: FDR < 0.05, $\log_2$FC $> \pm 1.2$). The genes upregulated were enriched in different pathways such as inflammation, defense and immune response possibly due to the inflammaging that occurs with age in the killifish (Fig. 4b–d). We further analyzed the data to find many genes related to the interferon signaling pathway to be enriched and upregulated in our gene list. In the upregulated list, we found genes associated with overall inflammation such as *cd68*, *aif1* and *pfn1*. The downregulated gene list included proliferation and development associated genes (Figs. 4b,e,f). Specifically, the downregulated genes included stemness and proliferation markers such as *kiaa0101*, *hes2*, *nfixb*. The full list of up- or down-regulated gene/isoforms is noted (Suppl. Table 1). The sample-wise expression levels of these key genes have been consolidated (Suppl. Figure 3). We noticed a similar response at the bulk level of our aged single-cell sequencing dataset confirming these findings[10].

Retained introns and skipped exons events recurrent upon aging. Alternative splicing is important for building the diversity of the transcriptome. We used Quant-seq expression data to study alternative splicing (AS) among the identified isoforms resulting from Iso-seq analyses. Using SUPPA2, we quantified diverse alternative splicing events that differ between the young versus aged transcriptomes. Local AS event inclusion levels are quantified by Proportion or Percentage Spliced in (PSI) levels which were calculated for all events that were enriched during aging. For the gene counts, we utilized a bulk RNA-seq differential aging-associated gene list. We found 29 significant events that were altered upon aging, specifically linked to neuronal genes such as *gap43* and *tubb5*, which were both downregulated. Other genes and their events that were upregulated included *rgs9b* and *tldc5* (Fig. 5a, b). The full list of altered events is available (Suppl. Table 2). We only found a small number of AS events significantly altered upon aging but it was interesting to note the type of AS event altered. More than half of the splicing changes were caused by Retained Intron (RI) and Skipped Exon (SE) events, and their ratios changed upon aging. The overall number of RI events were reduced within the dataset, while SE events increased upon aging (Fig. 5c). We further performed pathway analysis of the 29 altered genes and found changes in gap junction, ribosome, spliceosome, and phagosome (Fig. 5d). These changes are commonly found alterations upon aging[22]. Fold enrichment between 20–80 was noticed, although number of genes per pathway was low (between 1-2). The pathways are ordered by their significance values (-log10(FDR)).

While 29 events may appear modest, similar studies in other vertebrate models, particularly using equivalent stringency, report diverse numbers when focusing on specific brain regions or aging-related changes. For instance, during zebrafish embryogenesis, studies using comparable short- and long-read integration have identified ~2500 splicing changes although it does not take the aging factor into account[23]. This indicates that major changes such as those which occur during development seem to bring in significant changes in the splicing landscape. In the mouse hippocampus, exon skipping dominates alternative splicing in aging and accounts for the majority of differential events, typically in the range of 50–150, though many are filtered out due to low transcript support[24,25]. Human brain studies show hundreds of events, but these often include noise due to high biological variability but, rely on much larger sample sizes for statistical power[26]. Our focused design based on the telencephalon of *N. furzeri*, and the integration of long- and short-read data presents a stringent, accuracy-prioritized way. The filtered splicing events we retained are more likely to represent functionally relevant isoform switches, particularly in genes related to neurogenesis, RNA processing, and synaptic function, processes that also surfaced from the pathway analyses (Fig. 5d).

Code availability

All software used in this study is public and their parameters are clearly described in the Methods. If no detailed parameters were mentioned for the software, default parameters were used as suggested by the developer.

Data availability

The complete Isoseq assembly transcriptome and the predicted gene sequence files (main dataset) are available on Figshare[20]. All RNA-seq raw reads of killifish telencephalon were deposited on NCBI SRA under project accession number: PRJNA1194046/SRP549330[19].

Received: 12 December 2024; Accepted: 30 September 2025;
Published online: 17 November 2025

References

1. Cellerino, A. & Ori, A. What have we learned on aging from omics studies? *Semin. Cell Dev. Biol.* **70**, 177–189 (2017).
2. Perez-Gomez, A., Buxbaum, J. N. & Petrascheck, M. The aging transcriptome: read between the lines. *Curr. Opin. Neurobiol.* **63**, 170–175 (2020).
3. Reinar, W. B. *et al*. Teleost genomic repeat landscapes in light of diversification rates and ecology. *Mob. DNA* **14**, 14 (2023).
4. Reichwald, K. *et al*. Insights into Sex Chromosome Evolution and Aging from the Genome of a Short-Lived Fish. *Cell* **163**, 1527–1538 (2015).
5. Van houcke, J. *et al*. Aging impairs the essential contributions of non-glial progenitors to neurorepair in the dorsal telencephalon of the Killifish *Nothobranchius furzeri*. *Aging Cell* **20** (2021).
6. de Bakker, D. E. M. & Valenzano, D. R. Turquoise killifish: A natural model of age-dependent brain degeneration. *Ageing Res. Rev.* **90**, 102019 (2023).
7. Bergmans, S., Raes, L., Moons, L. & De Groef, L. A review on neurodegeneration in the fast-ageing killifish, the first animal model to study the natural occurrence of neuronal cell loss. *Ageing Res. Rev.* **91**, 102065 (2023).
8. Van houcke, J. *et al*. Modeling Neuroregeneration and Neurorepair in an Aging Context: The Power of a Teleost Model. *Front. Cell Dev. Biol.* **9** (2021).

9. Zhang, D. *et al*. Incomplete annotation has a disproportionate impact on our understanding of Mendelian and complex neurogenetic disorders. *Sci. Adv.* **6**, eaay8299 (2020).
10. Ayana, R. *et al*. Single-cell sequencing unveils the impact of aging on the progenitor cell diversity in the telencephalon of the female killifish N. furzeri. *Aging Cell* **n/a**, e14251.
11. Valenzano, D. R. *et al*. The African Turquoise Killifish Genome Provides Insights into Evolution and Genetic Architecture of Lifespan. *Cell* **163**, 1539–1554 (2015).
12. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
13. ToFU. https://github.com/Magdoll/cDNA_Cupcake.
14. Tardaguila, M. *et al*. SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res.* **28**, 396–411 (2018).
15. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
16. Trincado, J. L. *et al*. SUPPA2: Fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol.* **19**, 40 (2018).
17. Ge, S. X., Jung, D. & Yao, R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics* **36**, 2628–2629 (2020).
18. Dobin, A. *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
19. *NCBI Sequence Read Archive*. http://identifiers.org/ncbi/insdc.sra:SRP549330 (2025).
20. Ayana, R. *et al*. Full-length transcriptome and isoform diversity of African turquoise killifish, Nothobranchius furzeri. https://doi.org/10.6084/m9.figshare.27969174.v3 (2024).
21. Nickless, A., Bailis, J. M. & You, Z. Control of gene expression through the nonsense-mediated RNA decay pathway. *Cell and Bioscience* **vol. 7** 26 at https://doi.org/10.1186/s13578-017-0153-7 (2017).
22. Van houcke, J. *et al*. A short dasatinib and quercetin treatment is sufficient to reinstate potent adult neuroregenesis in the aged killifish. *npj Regen. Med.* **8**, 31 (2023).
23. Liu, Z. *et al*. Temporal Dynamic Analysis of Alternative Splicing During Embryonic Development in Zebrafish. *Front. Cell Dev. Biol.* **10**, 879795 (2022).
24. Stilling, R. M. *et al*. De-regulation of gene expression and alternative splicing affects distinct cellular pathways in the aging hippocampus. *Front. Cell. Neurosci.* **8**, 373 (2014).
25. Mazin, P. *et al*. Widespread splicing changes in human brain development and aging. *Mol. Syst. Biol.* **9**, 633 (2013).
26. Raj, T. *et al*. Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility. *Nat. Genet.* **50**, 1584–1592 (2018).

Acknowledgements

The authors also would like to acknowledge the Genomics Core facility of KU Leuven for their technical help in sequencing and preliminary analysis of the data. We used Biorender for the illustrative figures presented in this paper.

Author contributions

R.A. was involved in conceptualization, experimental design, bioinformatics analysis, data deposition, data visualization, writing: draft, review and editing. T.K. was involved in bioinformatics analyses, data visualization, writing: review and editing. C.Z. was involved in bulk RNA-seq sample preparation, writing: review and editing. J.V.H. and V.M. were involved in sample preparation for long-read sequencing, writing: review and editing. G.E.M. was involved in both sequencing experiment preparations, writing: review and editing. L.A. and E.S. were involved in study supervision, conceptualization, experimental design, funding, writing: draft, review and editing. All authors approved the submission.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-06073-7.

Correspondence and requests for materials should be addressed to R.A., E.S. or L.A.