

OBSCN undergoes extensive alternative splicing during human cardiac and skeletal muscle development

Ali Oghabian^{1,2}, Per Harald Jonson^{1,3}, Swethaa Natraj Gayathri^{1,3}, Mridul Johari^{1,4}, Ella Nippala¹, David Gomez Andres⁵, Francina Munell⁵, Jessica Camacho Soriano⁶, Maria Angeles Sanchez Duran⁷, Juha Sinisalo⁸, Heli Tolppanen⁹, Johanna Tolva¹⁰, Peter Hackman^{1,3}, Marco Savarese^{1,3*}, Bjarne Udd^{1,3,5, 11*}

1 Folkhälsan Research Center, Helsinki, Finland

2 Research Program for Clinical and Molecular Metabolism, Faculty of Medicine, University of Helsinki, 00014 Helsinki, Finland.

3 Department of Medical Genetics, Medicum, University of Helsinki, Helsinki, Finland

4 Harry Perkins Institute of Medical Research, Centre for Medical Research, University of Western Australia, Nedlands, WA, Australia.

5 Pediatric Neuromuscular Unit. Child Neurology Department. Hospital Universitari Vall d'Hebron, Vall d'Hebron Research Institute (VHIR) ES.

6 Histology Department, Vall d'Hebron University Hospital ES.

7 Prenatal Diagnosis Department, Vall D'hebron University Hospital ES.

8 Maternal Fetal Medicine Unit, Department of Obstetrics, Universitat Autònoma de Barcelona, Hospital Vall D'hebron, Barcelona, Spain

9 Helsinki University Central Hospital, Helsinki, Finland

10 Transplantation Laboratory, Department of Pathology, University of Helsinki, Finland

11 Neuromuscular Research Center, Department of Neurology, Tampere University and University Hospital, Tampere, Finland

*Equal contribution **Corresponding author:** Ali Oghabian (ali.oghabian@helsinki.fi)

Additional File 1: Supplemental methods and Figures S1-4

Supplemental methods

Figure S1. The distinction of samples of different sexes and clinical diagnoses based on *OB-SCN* exon inclusion.

Figure S2. Heatmap plot of inclusion levels of the studied exons of *OBSCN*.

Figure S3. Inclusion levels of the studied exons of *OBSCN*.

Figure S4. Alternative 3' splicing.

Supplemental methods

In-house data

For prenatal analysis, a trained fetal pathologist collected fetal skeletal muscles (n=20) and fetal cardiac muscles (n=2) from 2 different fetuses, without muscle pathology, obtained from voluntary termination of pregnancy (TOP).

For postnatal analysis, we collected sample biopsies from an internal cohort of 44 individuals (Table 1).

RNA was extracted with the Qiagen RNeasy Plus Universal Mini Kit (Qiagen, Hilden, Germany) according to the instructions provided by the manufacturer. Total RNA-Seq libraries were prepared using the Illumina Ribo-Zero Plus rRNA Depletion Kit (Illumina, Palo Alto, CA, USA) at the Oxford Genomics Center, Wellcome Trust Institute, Oxford, United Kingdom and Novogene. Sequencing was performed using NovaSeq 6000 (Illumina), generating over 80 million 150bp-long reads per sample.

External data

In addition to our in-house data, we studied four samples from ENCODE with accession IDs: ENCBS067RNA (fetal skeletal muscle tissue, 19 weeks female), ENCBS068RNA (fetal skeletal muscle tissues, 22 weeks male), ENCBS055RNA (fetal heart tissue, 28 weeks female) and ENCBS056RNA (fetal heart tissue, 19 weeks female) [27,28]. Furthermore, for a view on *OBSCN* expression across different human tissues, we obtained data from the GTEx Portal (Analysis Release V8) and dbGaP accession number phs000424.v8.p2 on 09/13/23.

RNA-Seq read alignment

The paired RNA-Seq reads were mapped to the Human Genome (GRCh38.p13) using the splice-aware alignment software STAR (V2.7.7a) [29]. All parameters relevant to the alignment process were set to their default settings except for *--outSAMattributes* that was set to *ALL*, *--sjdbScore* to *1*, *--limitOutSJcollapsed* to *2000000*, and *--outMultimapperOrder* to

Random. Furthermore, for a more sensitive exon junction processing (which leads to increased read counts for the exon junctions), the software was run in 2-pass mode by setting the *--twopassMode* parameter to *Basic*. For the gene annotation, Gencode.v39 was used.

Splicing analysis and exon inclusion level estimation

The inclusion levels (*i.e.* PSI or Ψ values) of all unique exons in human genome, including those of *OBSCN* gene, were measured using the Intron Exon Retention Estimator (IntERESt) R/Bioconductor package (V1.26.1) [30]. IntERESt is a comprehensive RNA-Seq read summarization, differential intron retention and splicing analysis software. It supports methods to summarize the reads that span the introns, map to the introns/exons, skip the exons, *etc.* It also supports tools that measure suitable Ψ values and run statistical differential test for splicing analysis. The inclusion Ψ values were measured for every *OBSCN* exon. It is 100 times the fraction of the reads that span the flanking introns (*i.e.* IS), to the sum of this number with twice the number of reads that skip the exon (*i.e.* ES) (*formula 1*). The measured Ψ (in percentage scale) is a value in the range of [0,100]. If all reads support the inclusion of an exon (*i.e.* splicing of the spanning introns) then Ψ would equal to 100. On the contrary, if all reads skip the exon, Ψ would be zero. Note that we exclusively considered the paired reads, which both pairs were mapped within the genome coordinates of the *OBSCN* gene. This was to exclude any reads from the gene fusions.

$$\Psi_{IR} = 100 \times \frac{IS}{IS+2 \times ES} \quad (\text{formula 1})$$

The statistical significance of the increase or decrease of the inclusion levels of *OBSCN* exons (for all exons except for the first and last exons of the studied transcripts) was estimated using the *deseqInterest()* function [30], *i.e.* a *DESeq2* dependent differential splicing analysis function supported by IntERESt that assumes that the read counts follow a negative binomial distribution [31]. The differential exon inclusion analysis was performed genome-wide for all exons, however later the results for the *OBSCN* exons were extracted. The statistical test

compared variations in the inclusion of each exon relative to the genome-wide variation observed for the inclusion of the exons. The analysis was adjusted for possible biases introduced by the different sequencing batches by including this parameter as a covariate in the design model of the statistical tests. For PCA analysis, the *prcomp()* function of the *stats* package of R (V3.4.0) was used. For the heatmap plots, the *heatmap.2()* function supported by the *gplots* package (V3.1.3.1) was used, whilst the Euclidean distance was used for hierarchical clustering.

In addition to exon skipping/inclusion, we analyzed the inclusion levels of the alternative first and last exons. All exons that were annotated as the first/last exon of the studied isoforms were extracted using biomaRt [32]. For genes other than *OBSCN*, the analysis was restricted to isoforms with RefSeq ID annotations. This was to avoid analyzing many isoforms with incomplete annotated 5' or 3' ends. It is worth noting that the Ψ values of these exons cannot be estimated with the same method that was used for the other exons (*formula 1*) as they lack exon-skipping sequence reads. Instead, for these exons, we define Ψ values specific to the inclusion of alternative first (AF) and last (AL) exons (*formula 2*). These values can be measured for any first (and last) exon of an isoform for which at least one alternative first (and respectively last) exon is featured in other isoforms. For any gene with M known first exons, Ψ_{AF_n} is the number of reads that span the downstream intron of the n th exon relative to the sum of reads that span the downstream introns of all alternative first exons, in percentile scale (*formula 2*). A similar benchmark was used for the alternative last exons, with the exception that the reads spanning the upstream introns were taken into account to measure Ψ (*formula 2*). Furthermore, similar to the exon inclusion analysis, *DESeq2* was used to analyze the statistical significance of the Ψ changes genome-wide. We compared these values in the postnatal heart and postnatal muscle samples to fetal heart and fetal muscle samples, respectively.

All P-values were adjusted for multiple testing using the Benjamini-Hochberg method [33].

An FDR < 0.05 cutoff was used to extract the significant results.

$$\Psi_{AF_n} = 100 \times \frac{IS_n}{\sum_{m=1}^M IS_m} \text{ (formula 2)}$$

Finally, for the only two exons annotated with alternative 3' splicing of upstream intron, *i.e.* exons 122 and 123 (or 119 and 119a from the meta transcript) (*Table 2*), we used a method similar to that we used for alternative last exons (formula 2) that measures the Ψ values based on the reads that span the upstream introns. Differential gene expression analysis was performed using DESeq2. We also adjusted for the possible batch effects. The rank correlation of VST normalized expression of the splicing factors (whose IDs were extracted from other studies [34–36]) with the Ψ values was also measured using the Spearman method. The gene expression levels were also adjusted for batch effects using the ComBat function of the ‘sva’ R/Biocoductor package.

Real-time polymerase chain reaction (RT-qPCR) validation

RNA was extracted from two adult muscle samples and two fetal muscle samples using the Qiagen RNeasy Plus Universal Mini Kit (Qiagen, Hilden, Germany) and according to the instructions provided by the manufacturer. The cDNA synthesis was performed using SuperScript III Reverse Transcriptase (Invitrogen TM) and random primers, according to the protocol provided by the manufacturer. The UCSC In-Silico PCR tool and Primer3web v4.1.0 were used to design primers to target either exon-exon junctions or regions near the junctions. The RT-qPCR assays were performed using the iQ SYBR Green Supermix (BIO-RAD) and 25nM of each specific primer. Furthermore, three technical replicates were taken into consideration. For the normalization, 18S was used as the reference gene. The final results were calculated using the $\Delta\Delta C_t$ method and the relative quantification (RQ) values were plotted (*Fig. 4S-T*).

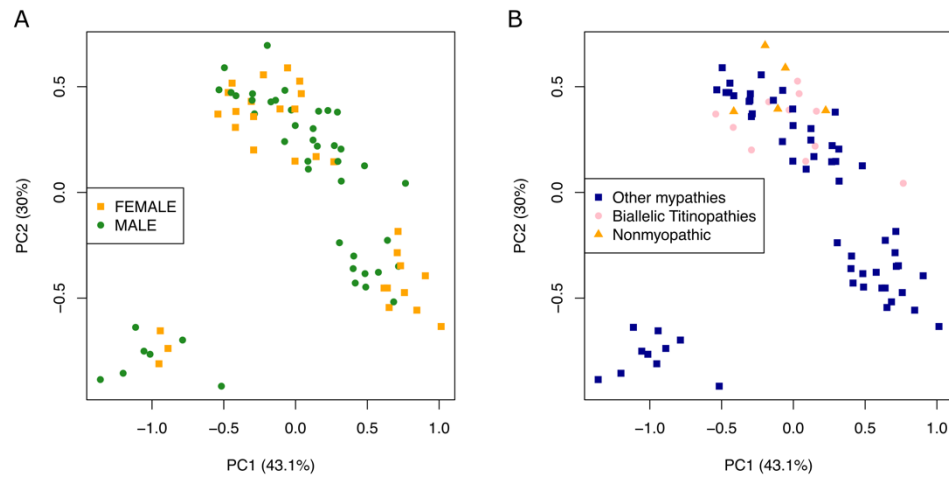


Figure S1. The distinction of samples of different sexes and clinical diagnoses based on *OBSCN* exon inclusion: The distinction of the studied sample groups based on the PC1 and PC2 dimensions. These dimensions were extracted from PCA analysis of the Percentage Spliced In (PSI) inclusion levels of exons of *OBSCN* across the studied samples. Each dot represents a sample and its colour and shape represent its corresponding group. The percentage of the variance within the data that is explained by the PC dimension is stated in parentheses in the axis labels. The dots have been coloured and distinguished based on 1) sex and 2) clinical diagnosis of the studied individuals.

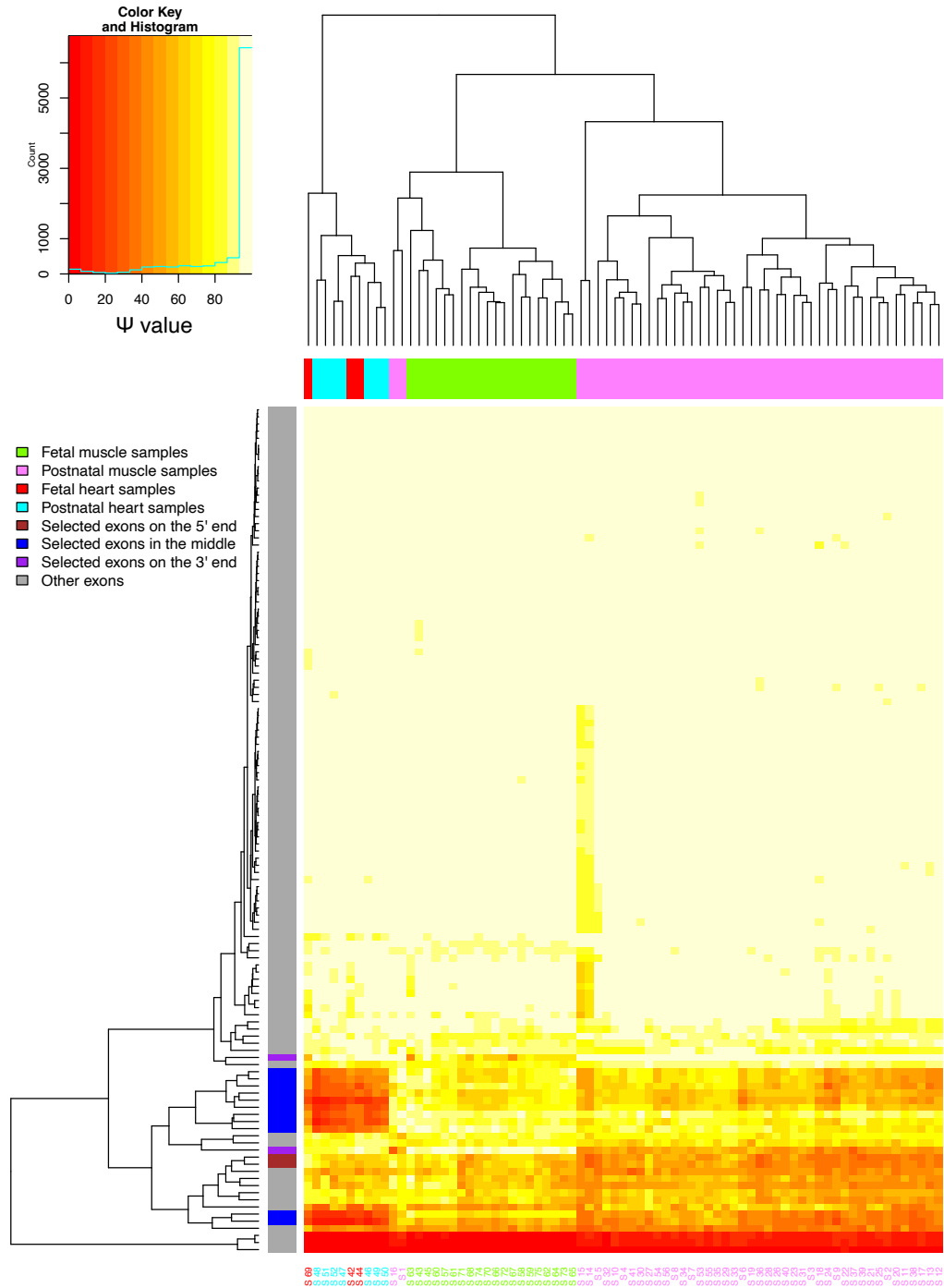


Figure S2. Heatmap plot of inclusion levels of the studied exons of *OBSCN*: The Percent-Spliced In (*i.e.* PSI or Ψ) inclusion level of the studied *OBSCN* exons in the studied samples is shown in the heatmap. The inclusion levels of the first and last exons of the transcripts are not shown as their measurements are not accurate. The colour key for the heatmap is shown in a box at the top left of the figure. The light-blue coloured line that goes through the

histogram shows the frequency of the Ψ values. The sample and the exon groups are labelled with distinctive colours. The exons are coloured based on how they were selected in the downstream analysis. The exons were selected if they were detected as significantly differentially spliced ($\text{FDR} < 0.05$) in at least three of the comparisons performed in the study. The Euclidean distance was used for hierarchical clustering.

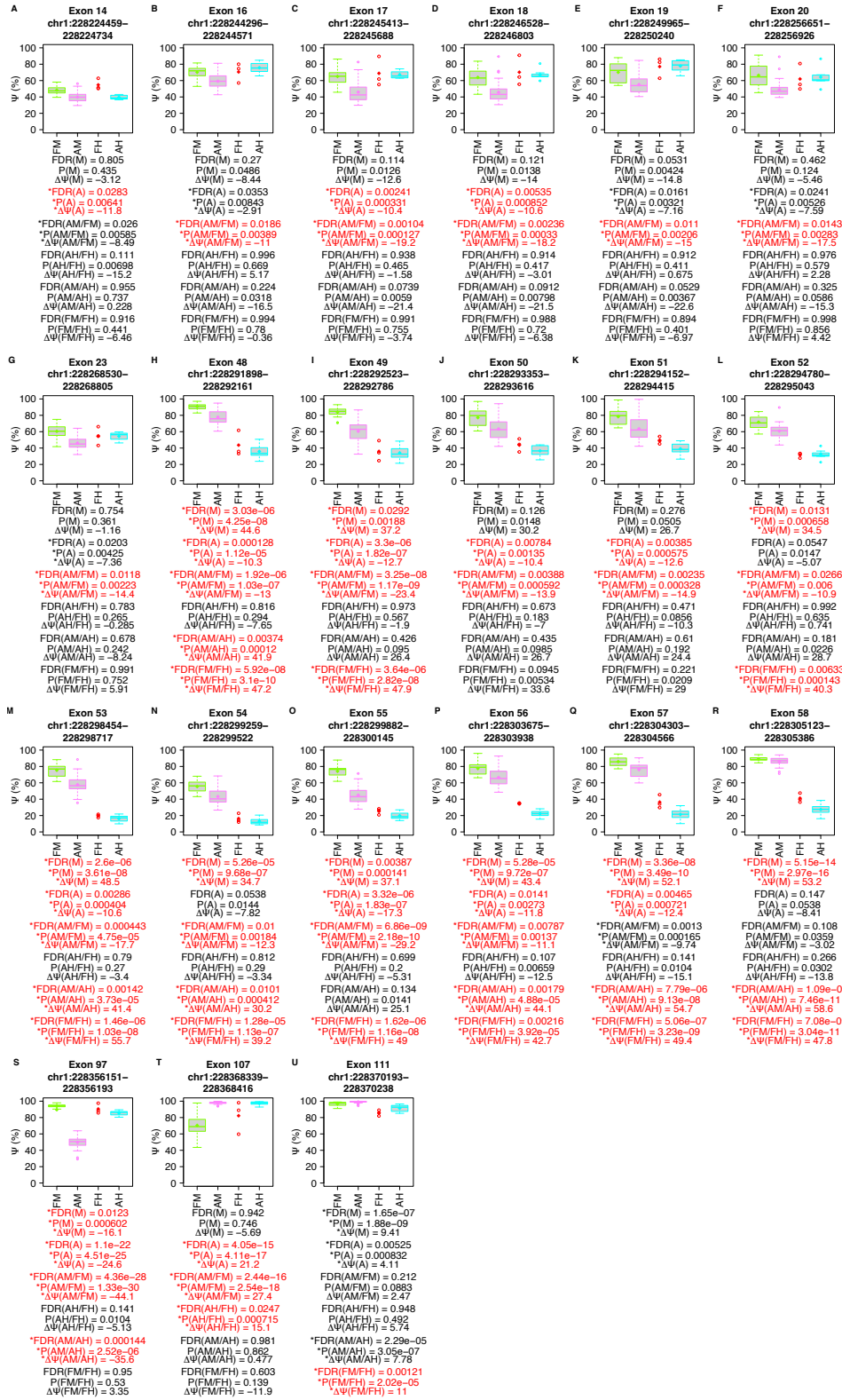


Figure S3. Inclusion levels of the studied exons of *OBSCN*: Boxplots illustrating the distribution of the Ψ levels of the studied *OBSCN* exons. The sample classes for which the Ψ values are shown are: postnatal muscles (AM), postnatal hearts (AH), fetal muscles (FM), and fetal hearts (FH). The FDR, P-value and $\Delta\Psi$ values for all the comparisons are listed below the box plots. The significant results (*i.e.* FDR < 0.05) are marked with * and coloured in red. Furthermore, the P-values and the FDR values of the first and last exons are shown with NA as no RNA-Seq reads skip these exons. The value for exon 126 is omitted, as its differences from exon 125 are minute (*i.e.* it extends on the 3' end by only 2 nucleotides). The comparisons include: muscle vs heart (M), postnatal vs fetal (A), postnatal muscle vs other samples (AM), postnatal heart vs other samples (AH), fetal muscle vs other samples (EM), and fetal heart vs other samples (FH). The box plots extend from the 25th to the 75th percentile, and the thick horizontal line represents the median. The boxplot whiskers show 1.5 times the interquartile range. The outliers are values higher and lower than the interquartile range.

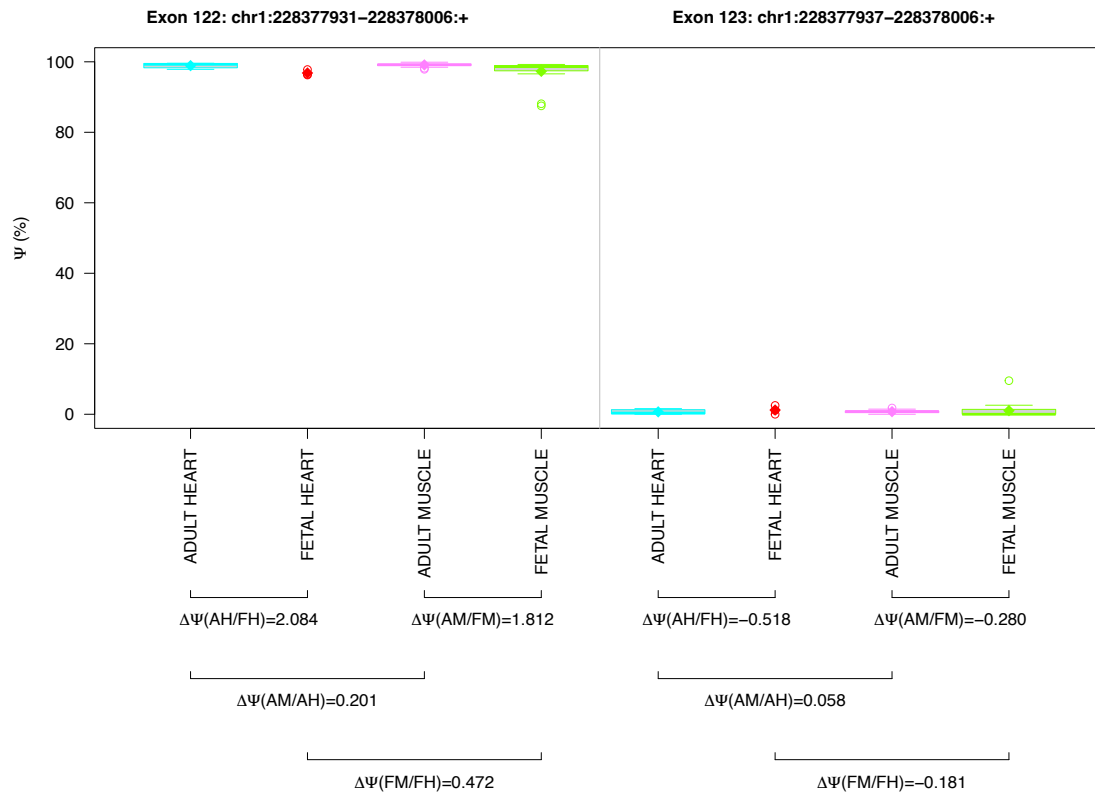


Figure S4. Alternative 3' splicing: Boxplots illustrating the distribution of the Ψ levels of the alternative 3' splicing. The headers show the coordinates of the affected exons. The sample classes for which the Ψ values are shown are: postnatal muscles (AM), postnatal hearts (AH), fetal muscles (FM), and fetal hearts (FH). The $\Delta\Psi$ values for two sets of comparisons are listed below the box plots: postnatal heart vs postnatal muscle (AH/AM), and fetal heart vs fetal muscle (EH/EM). The box plots extend from the 25th to the 75th percentile, and the thick horizontal line represents the median. The boxplot whiskers show 1.5 times the interquartile range. The outliers are values higher and lower than the interquartile range.