Ribosome profiling at isoform level reveals an evolutionary conserved impact of differential splicing on the proteome

Marina Reixachs-Solé[1], Jorge Ruiz-Orera[2], M Mar Alba[2,3], Eduardo Eyras[2,4,*],

[1]Pompeu Fabra University, E08003 Barcelona, Spain.

[2]IMIM - Hospital del Mar Medical Research Institute. E08003 Barcelona, Spain.

[3]Catalan Institution for Research and Advanced Studies. E08010 Barcelona, Spain

[4]The John Curtin School of Medical Research, Australian National University, Acton ACT 2601, Canberra, Australia

*correspondence to: eduardo.eyras@anu.edu.au

Abstract

The differential production of transcript isoforms from gene loci is a key cellular mechanism. Yet, its impact in protein production remains an open question. Here, we describe ORQAS (ORF quantification pipeline for alternative splicing) a new pipeline for the translation quantification of individual transcript isoforms using ribosome-protected mRNA fragments (Ribosome profiling). We found evidence of translation for 40-50% of the expressed transcript isoforms in human and mouse, with 53% of the expressed genes having more than one translated isoform in human, 33% in mouse. Differential analysis revealed that about 40% of the splicing changes at RNA level were concordant with changes in translation, with 21.7% of changes at RNA level and 17.8% at translational level conserved between human and mouse. Furthermore, orthologous cassette exons preserving the directionality of the change were found enriched in microexons in a comparison between glia and glioma, and were conserved between human and mouse. ORQAS leverages ribosome profiling to uncover a widespread and evolutionary conserved impact of differential splicing on the translation of isoforms and in particular, of microexon-containing ones. ORQAS is available at https://github.com/comprna/orqas

The alternative processing of transcribed genomic loci through transcript initiation, splicing, and polyadenylation, determine the repertoire of RNA molecules in cells [1]. Differential production of transcript isoforms, especially through the mechanism of alternative splicing, is crucial in multiple biological processes such as cell differentiation, acquisition of tissue-specific functions, and DNA repair [2–4], as well as in multiple pathologies [5–7]. Although analysis of RNA sequencing (RNA-seq) data from multiple samples has indicated a large diversity of transcript molecules [8], genes express mostly one single isoform in any given condition and this isoform may change across conditions [9,10].

Computational and in-vitro studies have provided evidence that a change in relative isoform abundances can lead to the production of protein variants that impact the network of protein-protein interactions in different contexts [11–14]. In contrast, quantitative proteomics of naturally occurring proteins has identified much fewer protein variants than those predicted with RNA sequencing [15,16]. Using state-of-the arts proteomics, it was recently shown that splicing changes at RNA level lead to changes in the sequence and abundance of proteins produced, although this was detectable for a limited number of transcripts [16]. The difficulty in establishing a correspondence between transcript and protein variation may be due to limitations in current proteomics technologies, but also to the stability and translation regulation of transcripts [17,18]. Despite the evidence about its functional relevance [3], it is still debated whether differential splicing leads to fundamentally different proteins and how widespread this might be [19–21]. Of particular interest are microexons, which can be as short as 3 nucleotides and carry out conserved neuronal-specific functions, and whose misregulation is linked to autism [22–24]. Despite their involvement in protein-protein interactions [23,25], the detection of protein variation associated to differential microexon inclusion using unbiased proteomics is currently not possible.

Sequencing of ribosome-protected RNA fragments, i.e. ribosome profiling, provides information on the messengers being translated in a cell. In particular, it allows the identification of multiple translated open reading frames (ORFs) in the same gene and the discovery of novel translated genes [26–29]. However, ribosome profiling studies have been mainly oriented to gene-level analysis [26,28,30]. Recently, reads from ribosome profiling have been mapped across the exon-exon junctions of alternative splicing events [31], suggesting that alternative splicing products may be engaged by ribosomes and potentially translated to produce different protein isoforms. A potential limitation of that approach is that ribosomal profiling reads also contain signals from native, non-ribosomal RNA-protein complexes [32]. As exon boundaries are profusely bound by RNA binding proteins and splicing factors [33], the mapping of ribosome reads to these regions is not necessarily indicative of active translation. Additionally, ribosome activity is associated to signal periodicity and uniformity along open reading frames [34], which has not yet been tested in relation to transcript isoforms and alternative splicing. Thus, the extent to which alternative splicing, and in particular microexon inclusion, leads to the translation of alternative ORFs remains largely unknown.

In this article, we describe a new method, ORQAS (ORF quantification pipeline for alternative splicing), to quantify translation abundance at individual transcript level from ribosome profiling taking into account Ribosome signal periodicity and uniformity per isoform. We validated the translation quantification of isoforms using independent data from polysomal fractions and proteomics. We further found a concordance between differential splicing and differential translation, and obtained evidence for the differential translation of microexons that is conserved between human and mouse. ORQAS provides a powerful strategy to study the impacts of differential RNA processing in translation.

Results

Translation Abundance estimation at isoform level from Ribo-seq

We developed a new method, ORQAS (ORF quantification pipeline for alternative splicing), for the estimation of isoform-specific translation abundance (Fig. 1a) (Methods). ORQAS quantifies the abundance of open reading frames (ORFs) in RNA space from RNA sequencing (RNA-seq) in transcript per million (TPM) units, and assigns ribosome sequencing (Ribo-seq) reads to the same ORFs using RiboMap [35]. After the assignment of Ribo-seq reads to isoform-specific ORFs, ORQAS calculates for each ORF two essential metrics to determine their potential translation: Uniformity, calculated as a proportion of the maximum entropy of the read distribution, and the 3nt periodicity along the ORF (Methods).

We analyzed with ORQAS Ribo-seq and matched RNA-seq data from human and mouse glia and glioma [30], mouse hippocampus [36], and mouse embryonic stem cells [37] (Supp. Table 1). To determine which values of uniformity and periodicity would be indicative of an isoform being translated, we selected as positive controls genes with a single annotated ORF and with evidence of protein expression in all 37 tissues recorded in the Human Protein Atlas (THPA) [38]. We considered translated those ORFs within the 90% of the periodicity and uniformity distribution of these positive controls (Fig. 1b) (Supp. Fig. 1). This produced a total of 20709-20785 translated ORFs in human, and 13,019-17,515 in mouse (Supp. Table 2). Interestingly, a large fraction of the expressed protein-coding genes had multiple translated isoforms: 52,3%-54,9% of the genes in human (Figs. 1c and 1d) and 29.1%-35.9% in mouse (Supp. Figure 2).

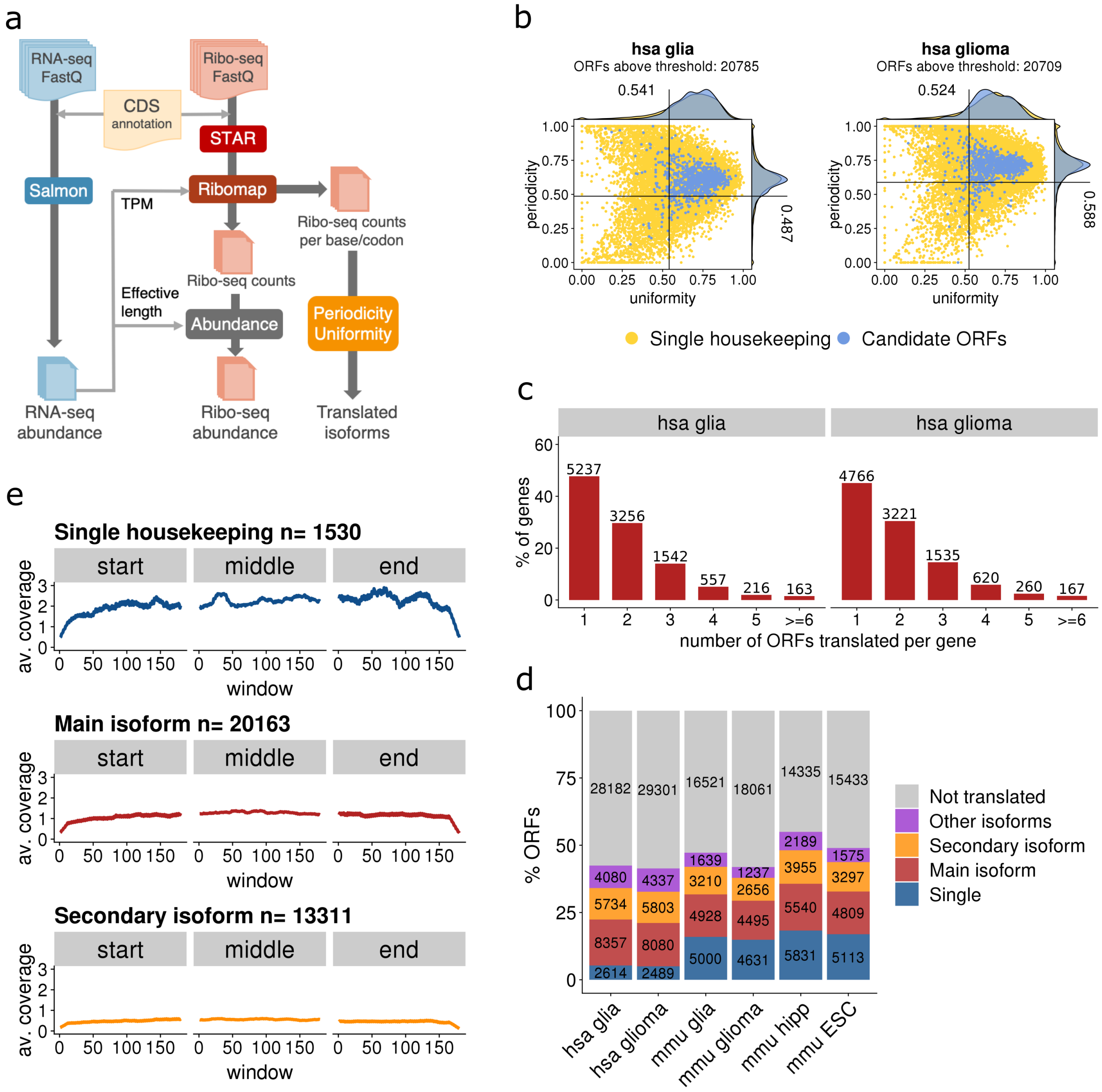

Figure 1. Estimation of translated isoforms. (a) The ORF quantification pipeline for alternative splicing (ORQAS) quantifies transcript abundances (ATXT) in transcripts per million (TPM) units from RNA-seq with Salmon [54] and in ORFs per million (OPM) units with Ribomap [35]. Coverage and periodicity are calculated for every ORFs with TPM > 0.1, and candidate translated isoforms are estimated by comparing to a set of single-ORF house keeping genes with protein expression evidence across 37 tissues. **(b)** Uniformity (x axis) versus periodicity (y axis) for all tested ORFs with RNA expression TPM > 0.1 and at least 10 Ribo-seq reads assigned (in yellow). Uniformity is measured as the percentage of maximum entropy and periodicity is measured in the first annotated frame. Single-ORF house keeping genes are indicated in blue. We show the data for human glia and glioma. Other samples are shown in Supp. Fig. 1. **(c)** Distribution of the number of different ORFs translated per gene in the human glia and glioma samples. Other samples are shown in Supp. Fig 2. **(d)** Number of ORFs predicted to be translated per sample, separated according to whether the ORF is encoded by: a single-ORF gene (Single), the most abundant isoform according to RNA-seq abundance in a gene with multiple isoforms (Main isoform), the second most abundant (Secondary isoform), or by any of the remaining isoforms in the abundance ranking (Other isoforms). Tested ORFs that are not predicted to be translated are depicted in gray (Not-translated). **(e)** Average density of Ribo-seq reads along ORFs in housekeeping singleton genes, in ORFs from the most

abundant isoform according to RNA-seq abundance in a gene with multiple isoforms, and in the second most abundant isoform.

Overall, the majority of translated isoforms correspond to either single-isoform genes or the isoform with the highest expression in a sample (main isoform) (Fig. 1d). However, from those genes with multiple isoforms expressed at the RNA level, 3,471-3,570 (52.6%-55.5%) of genes in human and 577-898 (27.6-34%) in mouse have an alternative isoform translated (Fig. 1d). From all translated isoforms, 47.3%-49.2% in human, and 28.3%-34.9% in mouse, correspond to alternative isoforms (secondary or other isoform, Fig. 1d). In genes with multiple isoforms, the main isoform showed the highest average Ribo-seq coverage compared to secondary isoforms, albeit not as high as for the single-ORF genes used as positive controls (Fig. 1e). As a quality control, we considered the proportion of isoforms with low or no RNA expression that fell inside our periodicity and uniformity cutoffs and found only 0.7-0.9% across the human samples and 0.1-1.5% in the mouse samples (Supp. Table 3).

Ribosome profiling discriminates translation abundance at isoform level

As an initial validation of the estimation of isoform-specific translation, we compared our predictions in human with immunohistochemistry data available from The Human Protein Atlas [38]. We observed that genes with translated isoforms are more frequently validated at all levels of protein expression (Fig. 2a). Furthermore, the majority (96%) of genes with translated ORFs show some evidence of protein expression using a combination of protein features (Fig. 2b). To further validate our approach, we compared the translated isoforms predicted with ORQAS with the sequencing of RNA from polysomal fractions from the same human neuronal and embryonic stem cell samples [39]. ORQAS predicted 27,552 translated isoforms in stem cells, and 25,034 in neurons (Supp. Fig. 3). We found that translated isoforms were enriched in polysomal fractions, whereas isoforms with RNA expression but not predicted to be translated with ORQAS were enriched in monosomal fractions (Fig. 2c), providing further support to our predictions. This is also consistent with a small proportion of our predicted translated isoforms to be associated with NMD targets, which are generally associated with monosomes [40].

Cross-species conservation is a strong indicator of stable protein production [41]. We thus decided to test the conservation of our translated isoforms in human and mouse, using glia and glioma samples available for both species. To this end, we used an optimization method to determine the human-mouse protein isoform pairs most likely to be functional orthologs (Methods) (Fig. 2d). From 15824 human-mouse 1-to-1 gene orthologs, we identified 18574 human-mouse protein isoform pairs representing potential functional orthologs. Moreover, 7,112 (64%) of the 1-to-1 gene orthologs had more than one orthologous isoform pair. We found that orthologous isoform pairs were significantly enriched in translated isoforms in both species (p-value < 2.2e-16 in all datasets) (Fig. 2e), providing further support for our predictions.

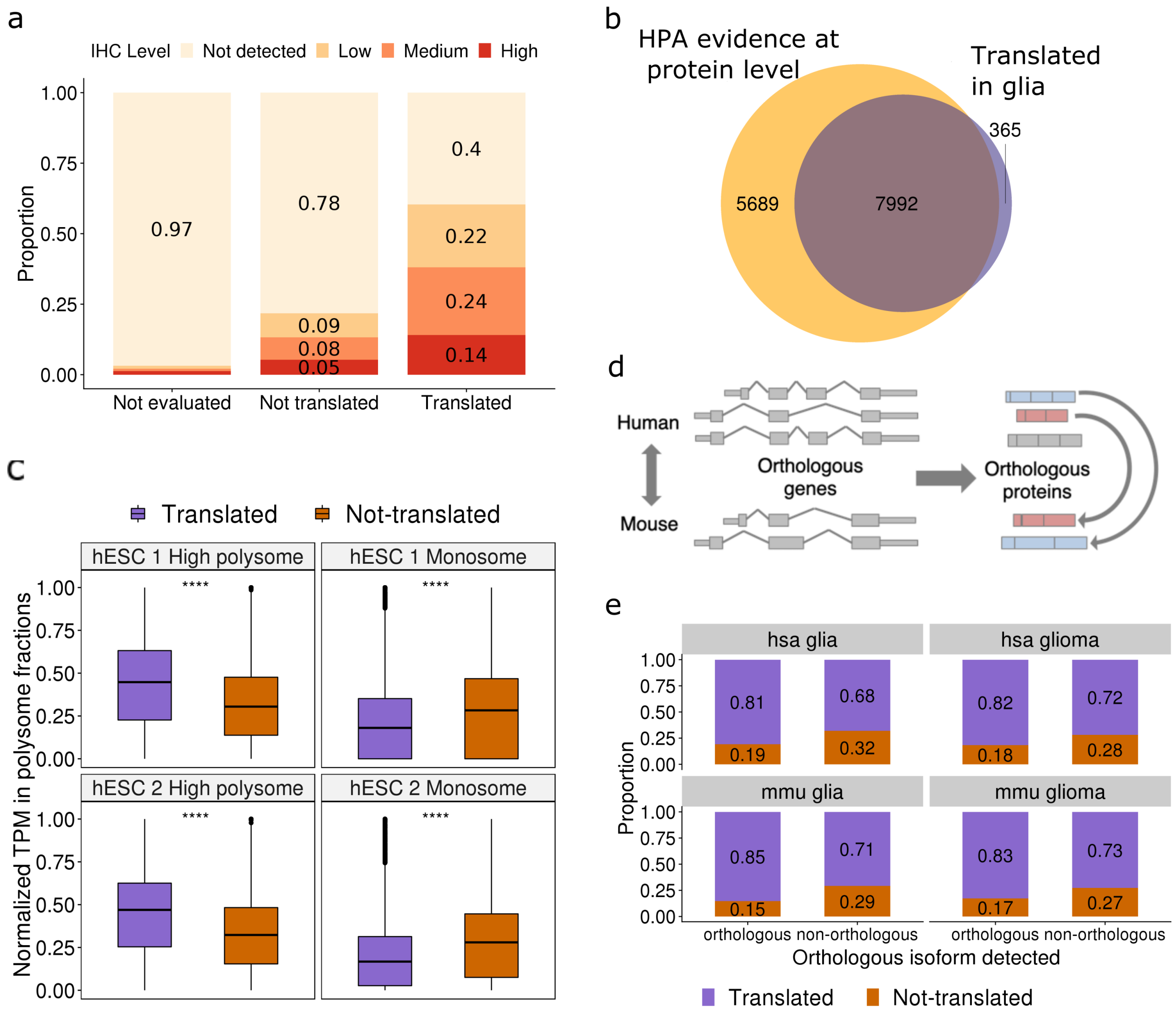

Figure 2. Validation of predictions. (a) For the candidate translated isoforms (translated), the cases that did not pass the threshold of uniformity and periodicity (not-translated), and those without enough read data to be tested (not evaluated), the plot shows the proportion of cases in which the corresponding gene has evidence from immunohistochemistry, separated as high, low, and medium expression, from the Human Protein Atlas [38]. We also indicate the cases not detected in the immunohistochemistry experiments (not-detected). The comparison was made for the human glia Ribo-seq. Singletons (single-ORF genes) were not included. **(b)** The plot shows the number of genes with predicted translated ORFs that have evidence of protein expression in the Human Protein Atlas from a combination of features: Mass Spectrometry, Immunohistochemistry and Uniprot. Translation predictions correspond to human glia Ribo-seq. Singletons were not included. **(c)** We show the distribution of the relative abundance in high polysome (left panels) and monosome (right panels) fractions for translated isoforms and for isoforms with RNA expression (TPM>0.1) but predicted as not translated. The plot shows the results for the two replicates for monosome (replicate 1 p-value = 3.61e-82 and replicate 2 p-value = 1.60e-152) and high polysome fractions for embryonic stem cells (replicate 1 p-value = 1.09e-230 and replicate 2 p-value = 2.41e-253). The results for neuronal cells are given in Supp. Fig. 3. **(d)** Cross-species conservation of protein isoforms. Protein isoforms from a 1-to-1 orthologous gene pair are compared and

candidate orthologous pairs are assigned using an optimization approach (Methods). **(e)** For the set of ORFs encoding a human-mouse orthologous protein pair (orthologous) and for those encoding proteins without an orthologous pair in mouse (non-orthologous) we plot the percentage that are predicted to be translated (translated) and the ones with RNA expression (TPM>0.1) but predicted as not translated (not-translated). We show here the results for human glia (p-value = 1.41e-140 Fisher test) and glioma (p-value = 3.63e-85), and for mouse glia (p-value = 1.143e-130) and glioma (p-value = 7.462e-53), Other mouse samples are shown in Supp. Fig. 3.

To perform an additional validation of our findings, we considered isoform-specific regions (Fig. 3a). We identified sequences that are unique to a specific isoform, since evidence mapped to these regions can be unequivocally assigned to the isoform. Additionally, we defined isoform-specific ORFs as sequences shared between two isoforms but with a different frame in each isoform, since protein evidence mapped to it can be confidently assigned to a specific ORF. Both region types in translated isoforms showed a higher density of reads per nucleotide compared with other isoforms (Fig. 3b) (Supp, Fig. 4a).

We further used peptides from Mass Spectrometry (MS) experiments [42] to match our predictions from Ribo-seq from the same tissue type (Methods). Overall we validated 86%-87% of translated single-ORF genes. Validation rate decreased with region length (Fig. 3c), as expected for MS experiments [41].

Additionally, since ORQAS quantification is performed for the entire ORF and not looking at specific regions within the ORF, we used the raw read data to validate the unique sequence regions. In total, 91-97% of unique sequence regions of length >200nt harbored uniquely mapping Ribo-seq reads (Fig. 3d) (Supp Fig. 4b), and 87-89% unique ORF regions of length > 200nt contained P-sites predicted from the mapped reads (Supp, Fig. 4c). Overall, we were able to validate 56-80% of the isoform-specific sequence regions tested and 48%-73% of the isoform-specific ORFs tested.

In summary, from all the protein-coding transcript isoforms considered from the annotation (84,024 in human and 48,928 in mouse), 58-59% in human and 63-65% in mouse showed RNA expression > 0.1 TPM (Supp. Table 3). From these expressed isoforms, about 40% in human, 41-54% in mouse, were predicted to be translated by ORQAS, and 23-43% were validated using independent data, including conservation (Fig. 3e). Furthermore, about 10% of all the annotated alternative isoforms in human and mouse had evidence of translation and these represented 60% of all translated isoforms (Fig. 3f) (Supp Table 4). Our analyses thus indicate that, although they are a small fraction of all expressed transcripts, alternative transcript isoforms are often translated into protein.

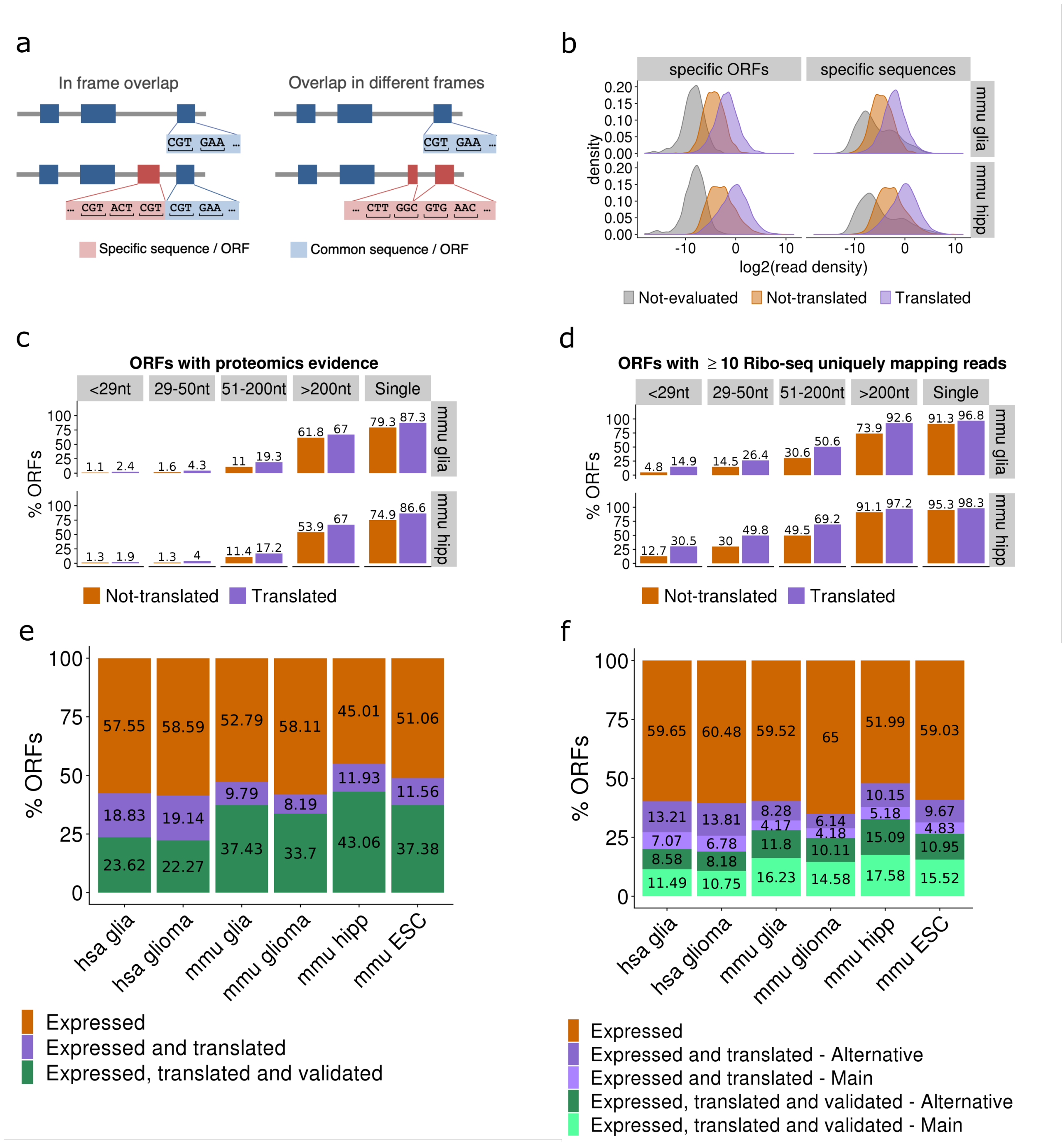

Figure 3. Validation with Isoform-specific regions. (a) Isoform-specific sequence regions (left panel) defined as the parts of an isoform ORF that are not present in any other isoform from the same gene (left panel). Isoform-specific ORFs (right panel) are defined AS a region shared between two isoforms, but with a different frame in each isoform. **(b)** For the mouse samples of glia (mmu glia) and hippocampus (mmu hipp) we show the density of Ribo-seq reads per nucleotide over the isoform-specific regions. The left panels show the counts per nucleotide based on the estimated P-site positions over region length in $\log_2$ scale for isoform-specific ORFs. The right panels show the uniquely mapping read-count over region length in $\log_2$ scale for isoform-specific sequences. Distributions are given for predicted translated isoforms, for isoforms that did not pass the threshold of uniformity and periodicity (not translated), and for the isoforms with low expression (TPM<0.1) (not evaluated). Other samples are shown in Supp Fig. 4. **(c)** Validation with mass spectrometry peptides. For the mouse samples of glia (mmu glia) and hippocampus (mmu hipp), the plot shows the percentage of ORF-specific regions with 1 or more peptides, separated according to region length. We show these results for both

types of regions: isoform-specific ORFs sequences and isoform-specific ORFs. Other samples are shown in Supp Fig. 4. **(d)** For the mouse samples of glia (mmu glia) and hippocampus (mmu hipp) the plot shows the percentage of regions with at least 10 uniquely mapping Ribo-seq reads in isoform-specific sequences over the total number of isoforms with an isoform-specific sequence defined according to the length of the region. Other samples are shown in Supp Fig. 4. **(e)** Proportion of isoforms expressed predicted to be translated and that have been validated using one or more sources of evidence: conservation, uniquely mapped Ribo-seq reads in specific sequences, counts per base in specific ORFs and peptides. For each sample, and for all ORFs with sufficient RNA-seq expression (TPM > 0.1), we show the proportion predicted to be translated from Ribo-seq reads and the proportion that were validated. The proportions are colored to indicate the fraction that is not included in the more restricted set: expressed > translated > validated. **(f)** For each sample, and for all ORFs with sufficient RNA-seq expression (TPM > 0.1), we show the proportion of main isoforms and alternative isoforms predicted to be translated from Ribo-seq reads and the proportion that were validated. These plots do not include the genes with a single protein-coding isoform.

Conserved impact of differential splicing on translation

Differential splicing is often assumed to lead to a measurable difference in protein production. However, at genome scale, this has only been shown for a limited number of cases [16]. We addressed this question using our more sensitive approach based on Ribo-seq. We used SUPPA [43,44] to obtain 37,676 alternative splicing events in human and 17,339 in mouse that covered protein coding regions (Methods). Using the same SUPPA engine to convert isoform abundances to event inclusion values [43,44], we estimated the proportion of translation abundance, relative abundance (RA), explained by a particular alternative splicing event, using the isoform translation abundances (Fig. 4a). Accordingly, in analogy to a relative inclusion change (ΔPSI) in RNA space, we were able to measure the relative differences in ribosome space in relation to the inclusion or exclusion of particular alternative exons, or ΔRA.

Comparing the glia and glioma samples in human, we found 856 events with a significant change in RNA splicing ($|\Delta PSI|>0.1$ and $p\text{-value}<0.05$), and 590 events with significant differential translation ($|\Delta RA|>0.1$ and $p\text{-value}<0.05$), with a significant overlap of 363 events between them (Jaccard index, z-score=89.386 comparing to the Jaccard index distribution of the overlaps from subsample sets of the same size) (Fig. 4b). Similarly, in mouse we found an overlap of 179 events (Jaccard index z-score=65.326), between 471 events with a significant change in RNA splicing ($|\Delta PSI|>0.1$ and $p\text{-value}<0.05$) and 240 with significant change in translation ($|\Delta RA|>0.1$ and $p\text{-value}<0.05$) (Supp. Fig. 5a). Furthermore, considering the direction of change from all events in RNA and ribosome space, the concordance of the change was found to be significant for human (Pearson R=0.9904, p-value = 5.309e-312) and mouse (Pearson R=0.9937, p-value = 2.113e-170); and in particular for the events that were significant in both tests (Fig. 4c) (Supp. Fig. 5b).

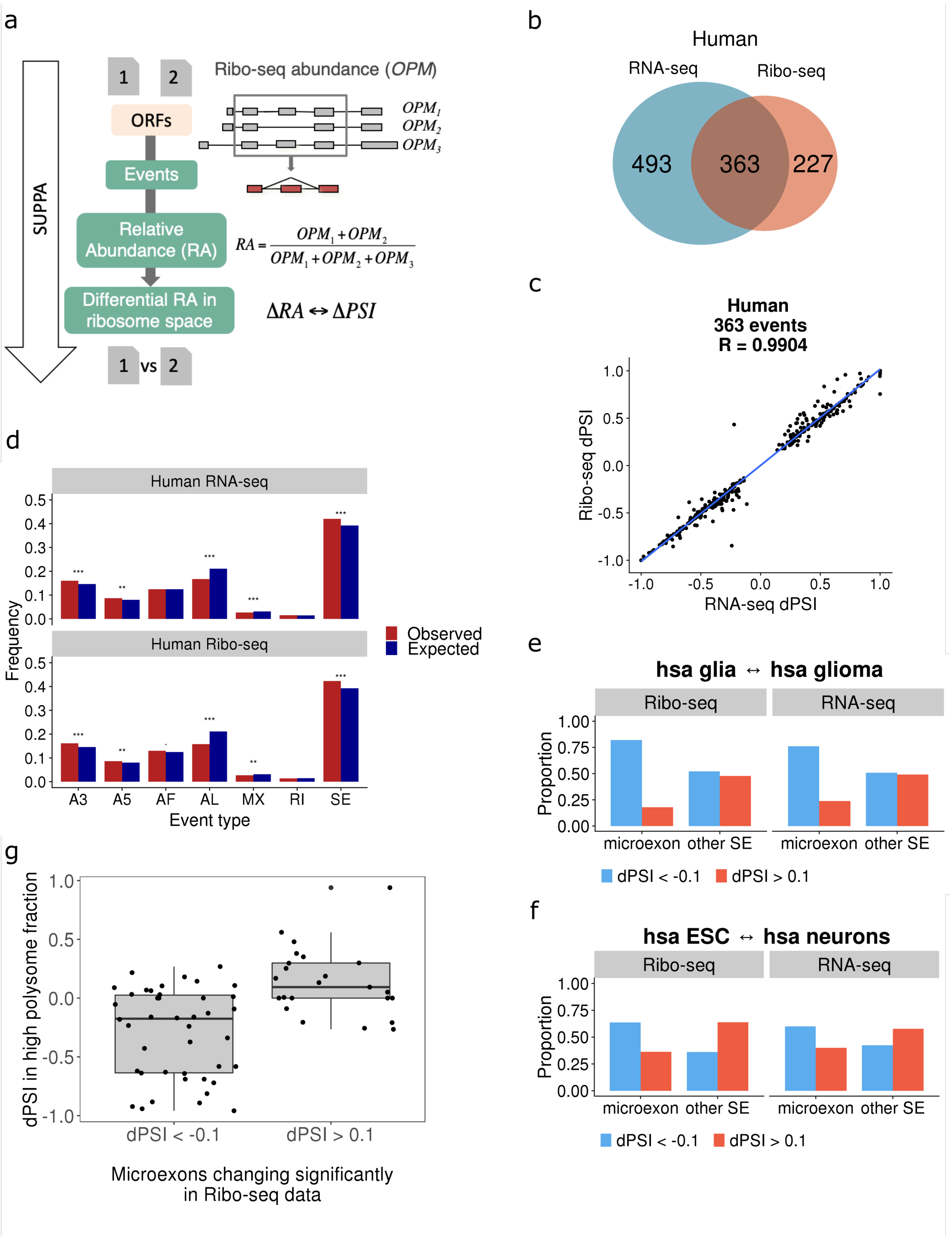

Figure 4. Differential translation linked to differential splicing. (a) Description of how SUPPA was used to calculate the differential inclusion of events in ribosome space. The abundance of open reading frames (ORFs) is calculated in ORFs per million (OPM) units. OPMs are transferred to events using SUPPA definition of events, using only exon-intron structures overlapping ORFs. For each event a relative abundance (RA) is obtained, analogous to a PSI. **(b)** Overlap of events changing significantly (dPSI > 0.1 and p-value < 0.05) with RNA-seq and with Ribo-seq for human. **(c)** Correlation

of changes in splicing and translation in events in human. **(d)** We show the proportions of events calculated in RNA space (upper panel) or in ribosome space (lower panel). In red we show the proportion of alternative splicing events calculated with SUPPA from the annotated coding regions in transcripts in human, whereas in blue we show the events that show a significant change using RNA-seq from mouse glia and glioma. Even types are: alternative 3'ss (A3) and 5'ss (A5), alternative first (AF) and last (AL) exon, mutually exclusive (MX) exon, retained introns (RI) and skipping exon (SE). There is significant enrichment for SE events for RNA (Fisher's test p-value = 3.90e-06) and Ribo-seq (p-value = 7.46e-07), for A3 events for RNA (p-value = 1.02e-05) and Ribo-seq (6.22e-06), for A5 events for RNA (4.55e-03) and Ribo-seq (8.73e-03); and significant depletion for AL events for RNA (1.73e-32) and Ribo-seq (2.43e-43) and for MX events for RNA (6.06e-04) and Ribo-seq (1.83e-03). These proportions for mouse are shown in Supp. Fig. 4. **(e)** Enrichment of microexons with an impact in RNA splicing and ORF translation in human from the comparison of glia and glioma samples. In the figure, dPSI is used to indicate the difference in relative abundance in both RNA and Ribosome spaces. **(f)** Enrichment of microexons with an impact in RNA splicing and ORF translation in human from the comparison of Embrionic Stem Cells (ESC) and neuronal samples. As before, dPSI indicates the difference in relative abundance in both RNA and Ribosome spaces. **(g)** Difference in high polysome fraction, measured as dPSI, between neuronal samples and ESCs (y axis) for microexons with a significant change in Ribosome space. As before, dPSI indicates the difference in relative abundance in Ribosome space.

We further observed a similar proportion of event types changing significantly in RNA and ribosome space, with an enrichment of exon skipping events in human (Fig. 4d) and mouse (Supp. Fig. 5c). In particular, microexons, defined to be of length ≤51nt [45], were enriched in the events changing between glia and glioma in both human (p-values 1.382e-12 for RNA-seq and 5.602e-10 for Ribo-seq) (Fig. 4e) and mouse (p-values 6.386e-16 for RNA-seq and 3.446e-06 for Ribo-seq) (Supp. Fig. 5d). We repeated the same analysis using data from human neuronal differentiation [39] and found that microexons were also enriched in the comparison between embryonic stem cells and neuronal cells in human for RNA splicing and translation changes (p-values 8.435e-06 for RNA-seq and 6.597e-05 for Ribo-seq) (Fig. 4f). Furthermore, using RNA sequencing from polysome fractions from the same stem cell and neuronal samples we were able to validate the change in inclusion patterns of microexons under the same conditions (Fig. 4g). These results provide evidence that differential splicing leads to a qualitative and quantitative change in the proteins produced from a gene locus. Our results are also consistent with a functional relevance of the inclusion of microexons in protein-coding transcripts in neuronal differentiation and their inclusion loss in brain-related disorders [22,23].

To further test the relevance of our findings, we considered a set of 1,487 alternative exons conserved between human and mouse (Fig. 5a). A high proportion of them changed in the same direction between glia and glioma (66% in RNA-seq and 78% in Ribo-seq) (Fig. 5b). Moreover, we observed that microexons were enriched in these concordant changes in both species, with a general trend towards less inclusion in glioma (p-value 5.389e-05 in RNA-seq and 5.521e-4 for Ribo-seq) (Fig. 5c). Among the microexons with a differential pattern of splicing and translation, we identified one in the gene *GOPC*, which was linked before to glioblastoma [46], and one in the gene

CERS6 (Fig. 5d). To test further the potential relevance of the identified microexons with conserved differential pattern, we calculated their RNA splicing inclusion patterns across other normal and tumor brain samples. In particular, we analyzed samples from glioblastoma multiforme (GBM) from TCGA [47], Neuroblastoma (NB) from TARGET [48] (Fig 5e), and samples from cortex and hippocampus from GTEX [49]. Microexons with a conserved impact on translation recapitulate the pattern of decreased inclusion in GBM compared with the postmortem normal brain regions (Fig. 5e). Differentially translated microexons may explain tissue differentiation as well as tumor specific properties, as they differentiate tissues and tumor types (Supp. Fig. 5e), whereas conserved microexons appear to be more representative of the tissue differentiation (Supp. Fig. 5f).

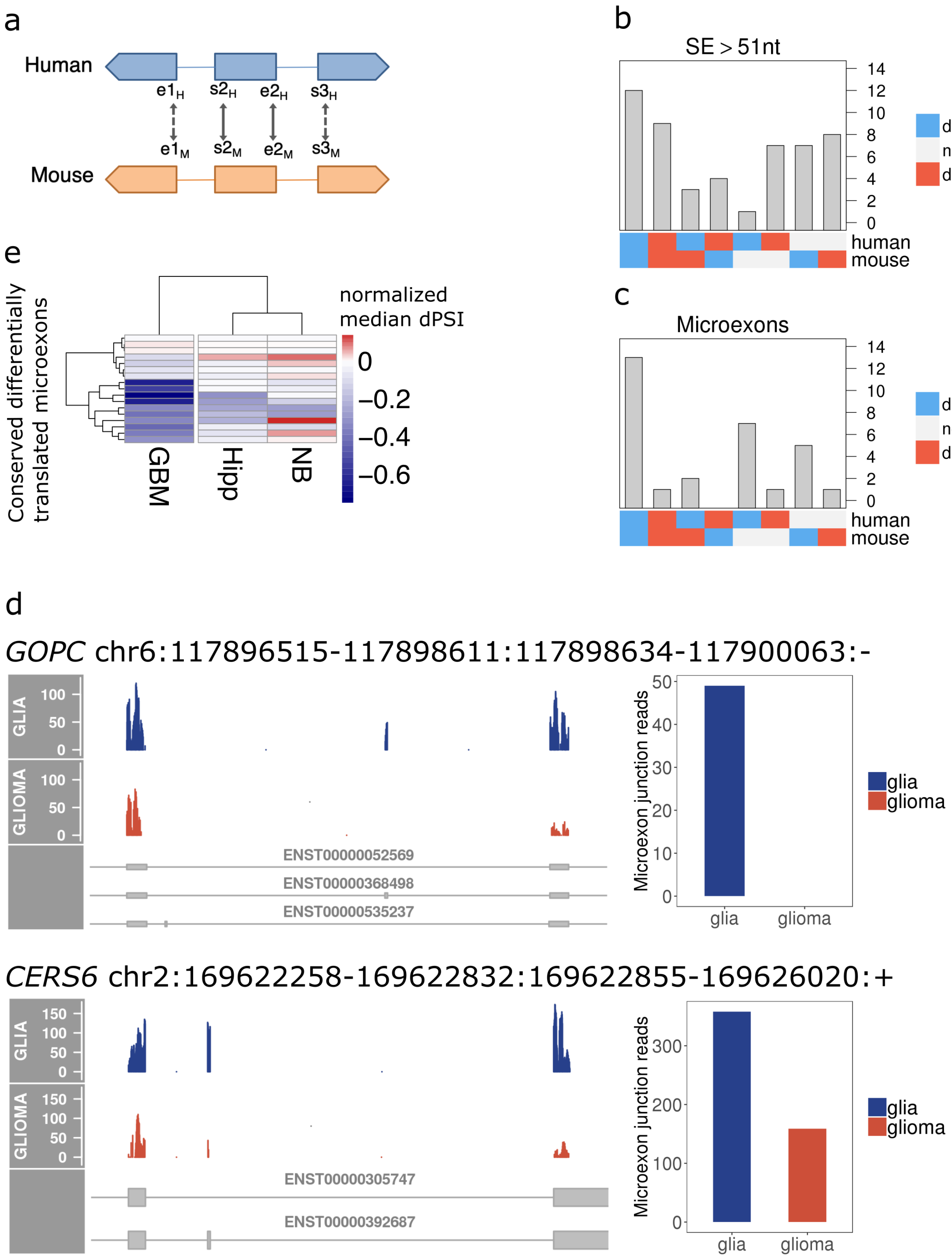

Figure 5. A conserved program of differential RNA usage and translation. (a) Conserved alternative splicing events were obtained by mapping with LiftOver the coordinates of the alternative exon (s2, e2), and the internal coordinates of the flanking exons (e1, s3). We considered those alternative exons as conserved if at least (s2, e2) were conserved. **(b)** Directionality of the changes in conserved alternative exons longer than 51nt in Ribosome space. As before, dPSI indicates here the difference in relative abundance in both RNA and Ribosome spaces. **(c)** Directionality of the changes in evolutionary conserved alternative microexons (≤51nt) in Ribosome space. As before, dPSI indicates here the difference in relative abundance in both RNA and Ribosome spaces. **(d)** Examples of microexons that change

significantly in RNA and ribosome space between glia and glioma. For the genes *GOPC* and *CERS6* we show Ribo-seq reads mapping to the microexon region and its flanking exons (left panels) and the number of Ribo-seq reads crossing the microexon junctions (right panels) in both glia (in blue) and glioma (in orange). **(e)** Patterns of inclusion of conserved differentially translated microexons in normal hippocampus (Hipp) samples from GTEX, gliblastoma multiforme (GBM) from TCGA, and neuroblastoma (NB) from TARGET. The heatmap shows the difference of the median PSI with respect to normal brain cortex tissue from GTEX.

Discussion

We have described a new method, ORQAS, to obtain abundance estimates at isoform level in ribosome space (https://github.com/comprna/orqas). ORQAS allows the identification of multiple protein products from a gene and the study of differential translation associated to alternative splicing and differential transcript usage between conditions. Our approach presents several novelties with respect to previous analyses [31,35,39] : *i)* we calculated the periodicity and uniformity for each isoform; *ii)* we validated our predictions using both sequence and ORF specific regions in isoforms, regardless whether these regions could be encoded into an standard alternative splicing event; and *iii)* we provided an isoform quantification in ribosome space that can be reused with other tools, like SUPPA. Additionally, ORQAS uses RNA-seq quantification to guide the isoform abundance estimation in ribosome space, unlike previous approaches that used directly Ribo-seq reads to quantify isoforms, which presents important limitations [39].

We estimated that in total about 40-50% of the protein coding isoforms with RNA expression showed some evidence of translation, and that around 20,700 proteins are produced in human and 13,000-17,500 in mouse in the tested conditions. Additionally, about 5,700-5,800 genes in human, 2,600-3,900 in mouse, produce more than one protein in those conditions. These estimates are far from what is generally predicted from RNA expression only [8]. This may be explained by the limited coverage of Ribo-seq reads, but may be also due to the fact that RNA-seq artificially amplifies fragments of unproductive RNAs leading to many false positives. Nonetheless, our data indicates that many more ORFs are translated in a given sample than what is detectable by current proteomics methods and the number of protein products are close to estimates using a combination of proteomics and sequence conservation [41]. Importantly, we found that multiple ORFs are translated from the same gene and at different abundances across conditions.

Around 40% of the events detected with differential RNA splicing showed consistent measurable changes in Ribo-seq in the same direction, which supports the notion that changes in RNA processing of genes have a widespread impact in the translation of ORFs from a gene. In particular, we found that a pattern of decreased inclusion of microexons in glioma with respect to normal brain samples is recapitulated in translation, providing *in vivo* evidence that the splicing changes in microexons have an impact in protein production. Microexon inclusion is a hallmark of neuronal differentiation [22,23,44], and glia partly recapitulates the pattern of microexon inclusion found in neurons [23]. The decreased inclusion of microexons observed in glioma suggests a dedifferentiation pattern as has been described before for other tumors [50], but could also be indicative of a difference in the content of neuronal cells in the samples compared. In either case, the evolutionary conservation of the change at RNA expression and protein production indicates a conserved functional program between the glia and glioma samples.

Our capacity to predict RNA splicing variations from RNA-seq data currently exceeds our power to evaluate the significance of those events regarding protein production with current proteomics technologies [51]. Despite this limitation, mass spectrometry can show for a small number of cases that splicing changes impact the abundance of proteins produced by a gene [16]. Our findings are in agreement with these results, and moreover overcome current limitations to determine genome-wide impacts of RNA processing changes on protein production. Furthermore, our analyses indicate that for the majority of genes producing multiple protein isoforms, these do not vary in more than 25% of the length of the most highly expressed isoform, suggesting that for most part, the functional impacts from alternative splicing are mediated by slight modifications in the protein sequences [25], rather than through the production of essentially different proteins. In summary, ORQAS leverages ribosome profiling to provide a genome-wide coverage of genes and transcript isoforms and allow a more effective testing of the impacts of splicing in protein production, as well as the identification and validation of multiple proteins from the same gene locus.

Online Methods

Pre-processing of RNA-seq and Ribo-seq datasets

RNA-seq and Ribo-seq datasets were downloaded from Gene Expression Omnibus (GEO) for the following samples: normal glia and glioma from human and mouse (GSE51424) [30], mouse hippocampus (GSE72064) [36], mouse embryonic stem cells (GSE89011) [37], and three steps of forebrain neuronal differentiation in human (GSE100007) [39]. Adapters in RNA-seq and Ribo-seq datasets were trimmed using cutadapt v.1.12 with additional quality filters (-hq = 30 -lq = 10) [52]. We further discarded reads that mapped to annotated rRNAs and tRNAs. Remaining reads in RNA-seq and Ribo-seq datasets were filtered by length (>= 26 nucleotides).

Quantification of transcripts and open reading frames

We used the Ensembl annotation v85 for human (hg19) and mouse (mm10) removing pseudogenes, short isoforms (< 200 nt) and annotated isoforms with incomplete 5' or 3' regions. For the analysis of RNA-seq data we used Salmon v0.7.2 (Patro et al. 2017) to quantify transcript abundances in transcripts per million (TPM) units using the annotation of unique open reading frames (ORFs). To quantify coding sequences (CDS) at the isoform level with the Ribo-seq data we applied a modified version of Ribomap (Wang et al. 2016). As default, Ribomap uses the RNA-seq reads aligned to the transcriptome sequences with STAR [53]. Instead, we provided a direct quantification of the ORFs with RNA-seq using Salmon, to be used as priors by RiboMap. We calculated the translation abundances of each ORF based on Ribo-seq in ORFs Per Million (OPM) units, analogously to the TPM units:

$$OPM_i = 10^6 \frac{n_i / l_i}{\sum_j n_j / l_j}$$

where n_i is the estimated Ribo-seq counts in ORF i and l_i is the effective length of the same ORF.

Identification of actively translated isoform coding sequences

We identified actively translated ORFs by calculating two parameters read periodicity and read uniformity [34]. The periodicity is based on the distribution of the reads in the annotated frame and the two alternative ones. In order to calculate the read periodicity, we previously computed the position of the P-site, corresponding to the tRNA binding-site in the ribosome complex. This was obtained by calculating the distance between annotated ATG start codons and the leftmost position covered by Ribo-Seq reads, for each read length, The uniformity was measured as the proportion of maximum entropy (PME) defined by the distribution of reads along the ORF:

$$H(X) = \sum_{i=1}^{n} \left(\frac{N_i}{N}\right) * log_2 \left(\frac{N_i}{N}\right)$$

$$PME = \frac{H(X)}{\max{(H)}}$$

Where *N* represents the total number of reads, *N_i* is the number of reads in each region *i* and *max(H)* is the entropy assuming that the reads are equally distributed across the ORF. The maximum value is 1, and indicates a completely even distribution of reads across codons. These values were obtained for each sample by pooling the replicates and we only considered ORFs with 10 or more assigned Ribo-seq reads, and with RNA-seq abundance TPM > 0.1.

Polysomal fraction analysis

We used RNA-seq from high polysomal, low polysomal and monosomal fractions from embryonic stem cells and neuronal cell culture in human (GSE100007) [39] to quantify isoforms with Salmon [54]. Only ORFs from protein-coding isoforms were used for quantification. For each isoform we calculated the polysomal relative abundance as before [17] by dividing the abundance in high polysomal fraction in TPM units, by the sum of abundances in (high and low) polysomes and monosomes.

Validation of isoform-specific regions

We defined two different types of isoform-specific regions that were analysed differently. Isoform-specific sequences are regions with a unique nucleotide sequence among the isoforms of the same gene. Isoform-specific ORFs are defined as regions that will give rise to different amino-acid sequences within the proteins of the same gene, either because of the presence of isoform-specific sequences or frame-shifted common sequences (Fig. 3A). According to the annotation, we identified 34553 isoforms containing isoform-specific sequences in human and 29447 in mouse and 44298 isoforms containing isoform-specific ORFs in human and 34329 in mouse. For the validation of isoform-specific sequences we considered uniquely mapping Ribo-seq reads from the STAR output falling entirely inside these regions or in the junction of the specific sequence with the common region. Read densities inside those regions where calculated as the total number of uniquely mapping reads in the region divided by the length of the isoform-specific sequence. The validation of isoform-specific ORFs instead was performed using the profiles of counts in each base of the ORF considering the expected position of the P-site. For isoform-specific ORFs the read densities where established as total number of counts in the region divided by the length in nucleotides of the isoform-specific ORFs.

Proteomics evidence in translated isoform coding sequences

We mined the proteomics database PRIDE (Vizcaino et al. 2016) to search for peptide matches to ORFs. We only considered peptide datasets from mouse corresponding to tissues analyzed in this study: brain (PRD000010, PXD000349, PXD001786), hippocampus (PRD000363, PXD000311, PXD001135), and embryonic cell lines (PRD000522). This corresponded to a total of 328,200 peptides. We searched for peptide matches in translated ORFs and only kept peptides that had one perfect match to an ORF and did not have a match with 0, 1 or 2 amino acid mismatches to any other annotated ORF isoform from the same or different genes.

Differential inclusion of events at RNA and translation level

We used SUPPA [43,44] to generate alternative splicing events defined from protein-coding transcripts and covering the annotated ORFs. The relative inclusion of an event was calculated with SUPPA in terms of the transcript abundances (in TPM units) calculated from RNA-seq and in terms of the ORF abundances (in OPM units) calculated from Ribo-seq. The test for significant differential inclusion of the events was applied in the same way for both cases by testing the difference between the observed change between conditions and the observed change between replicates, as described before [44].

Calculation of orthologous isoforms

We considered the set of 1-to-1 orthologous genes between human and mouse from Ensembl (v85) [55]. For each pair of orthologous genes we calculated all possible pairwise global alignments between the human and mouse proteins encoded by these genes using MUSCLE [56]. For each alignment we defined a score as the fraction of amino acid matches over the total length of the global alignment, and kept only protein pairs with score >= 0.8. From all the remaining protein pairs in each orthologous gene pair, we selected the best human-mouse protein pairs using a symmetric version of the stable marriage algorithm [57]

Acknowledgements

We acknowledge funding from the Spanish Government and FEDER with grants BFU2015-65235-P, BIO2017-85364-R and MDM-2014-0370, and by Catalan Government (AGAUR) with grant SGR2017-1020. MR-S was funded by an FI grant from the Catalan Government with reference 2018FI_B1_00126.

References

1. Brown, J. B. *et al.* Diversity and dynamics of the Drosophila transcriptome. *Nature* **512,** 393–9 (2014).
2. Fiszbein, A. & Kornblihtt, A. R. Alternative splicing switches: Important players in cell differentiation. *Bioessays* **39,** (2017).
3. Baralle, F. E. & Giudice, J. Alternative splicing as a regulator of development and tissue identity. *Nat. Rev. Mol. Cell Biol.* **18,** 437–451 (2017).
4. Shkreta, L. & Chabot, B. The RNA Splicing Response to DNA Damage. *Biomolecules* **5,** 2935–77 (2015).
5. Ward, A. J. & Cooper, T. A. The pathobiology of splicing. *Journal of Pathology* **220,** 152–163 (2010).
6. Singh, B. & Eyras, E. The role of alternative splicing in cancer. *Transcription* **8,** (2017).
7. Cummings, B. B. *et al.* Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci. Transl. Med.* **9,** (2017).
8. Pertea, M. *et al.* CHESS: a new human gene catalog curated from thousands of large-scale

RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biol.* **19,** 208 (2018).

9. Gonzàlez-Porta, M., Frankish, A., Rung, J., Harrow, J. & Brazma, A. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biol.* **14,** R70 (2013).
10. Sebestyén, E., Zawisza, M. & Eyras, E. Detection of recurrent alternative splicing switches in tumor samples reveals novel signatures of cancer. *Nucleic Acids Res.* **43,** (2015).
11. Buljan, M. *et al.* Tissue-Specific Splicing of Disordered Segments that Embed Binding Motifs Rewires Protein Interaction Networks. *Mol. Cell* **46,** 871–883 (2012).
12. Yang, X. *et al.* Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell* **164,** 805–17 (2016).
13. Climente-González, H., Porta-Pardo, E., Godzik, A. & Eyras, E. The Functional Impact of Alternative Splicing in Cancer. *Cell Rep.* **20,** 2215–2226 (2017).
14. Wojtowicz, W. M. *et al.* A vast repertoire of Dscam binding specificities arises from modular interactions of variable Ig domains. *Cell* **130,** 1134–45 (2007).
15. Ezkurdia, I. *et al.* Most highly expressed protein-coding genes have a single dominant isoform. *J. Proteome Res.* **14,** 1880–7 (2015).
16. Liu, Y. *et al.* Impact of Alternative Splicing on the Human Proteome. *Cell Rep.* **20,** 1229–1241 (2017).
17. Maslon, M. M., Heras, S. R., Bellora, N., Eyras, E. & Cáceres, J. F. The translational landscape of the splicing factor SRSF1 and its role in mitosis. *Elife* **2014,** (2014).
18. Braun, K. A. & Young, E. T. Coupling mRNA synthesis and decay. *Mol. Cell. Biol.* **34,** 4078–87 (2014).
19. Tress, M. L., Abascal, F. & Valencia, A. Alternative Splicing May Not Be the Key to Proteome Complexity. *Trends Biochem. Sci.* **42,** 98–110 (2017).
20. Blencowe, B. J. The Relationship between Alternative Splicing and Proteomic Complexity. *Trends Biochem. Sci.* **42,** 407–408 (2017).
21. Tress, M. L., Abascal, F. & Valencia, A. Most Alternative Isoforms Are Not Functionally Important. *Trends Biochem. Sci.* **42,** 408–410 (2017).
22. Raj, B. *et al.* A global regulatory mechanism for activating an exon network required for neurogenesis. *Mol. Cell* **56,** (2014).
23. Irimia, M. *et al.* A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* **159,** 1511–23 (2014).
24. Quesnel-Vallieres, M. *et al.* Misregulation of an Activity-Dependent Splicing Network as a Common Mechanism Underlying Article Misregulation of an Activity-Dependent Splicing Network as a Common Mechanism Underlying Autism Spectrum Disorders. *Mol. Cell* 1023–1034 (2016). doi:10.1016/j.molcel.2016.11.033

25. Ellis, J. D. *et al.* Tissue-specific alternative splicing remodels protein-protein interaction networks. *Mol. Cell* **46,** 884–92 (2012).

26. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324,** 218–23 (2009).

27. Michel, A. M. *et al.* Observation of dually decoded regions of the human genome using ribosome profiling data. *Genome Res.* **22,** 2219–29 (2012).

28. Ruiz-Orera, J., Messeguer, X., Subirana, J. A. & Alba, M. M. Long non-coding RNAs as a source of new peptides. *Elife* **3,** e03523 (2014).

29. Ruiz-Orera, J. *et al.* Origins of De Novo Genes in Human and Chimpanzee. *PLoS Genet.* **11,** e1005721 (2015).

30. Gonzalez, C. *et al.* Ribosome profiling reveals a cell-type-specific translational landscape in brain tumors. *J. Neurosci.* **34,** 10924–36 (2014).

31. Weatheritt, R. J., Sterne-Weiler, T. & Blencowe, B. J. The ribosome-engaged landscape of alternative splicing. *Nat. Struct. Mol. Biol.* **23,** 1117–1123 (2016).

32. Ji, Z., Song, R., Huang, H., Regev, A. & Struhl, K. Transcriptome-scale RNase-footprinting of RNA-protein complexes. *Nat. Biotechnol.* **34,** 410–3 (2016).

33. Witten, J. T. & Ule, J. Understanding splicing regulation through RNA splicing maps. *Trends Genet.* **27,** 89–97 (2011).

34. Ji, Z., Song, R., Regev, A. & Struhl, K. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *Elife* **4,** e08890 (2015).

35. Wang, H., McManus, J. & Kingsford, C. Isoform-level ribosome occupancy estimation guided by transcript abundance with Ribomap. *Bioinformatics* **32,** 1880–2 (2016).

36. Cho, J. *et al.* Multiple repressive mechanisms in the hippocampus during memory formation. *Science* **350,** 82–7 (2015).

37. Sugiyama, H. *et al.* Nat1 promotes translation of specific proteins that induce differentiation of mouse embryonic stem cells. *Proc. Natl. Acad. Sci. U. S. A.* **114,** 340–345 (2017).

38. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science* **347,** 1260419–1260419 (2015).

39. Blair, J. D., Hockemeyer, D., Doudna, J. A., Bateup, H. S. & Floor, S. N. Widespread Translational Remodeling during Human Neuronal Differentiation. *Cell Rep.* **21,** 2005–2016 (2017).

40. Kim, W. K. *et al.* mRNAs containing NMD-competent premature termination codons are stabilized and translated under UPF1 depletion. *Sci. Rep.* **7,** 15833 (2017).

41. Ezkurdia, I. *et al.* Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. *Hum. Mol. Genet.* **23,** 5866–78 (2014).

42. Vizcaíno, J. A. *et al.* 2016 update of the PRIDE database and its related tools. *Nucleic Acids*

Res. **44,** 11033 (2016).

43. Alamancos, G. P., Pagés, A., Trincado, J. L., Bellora, N. & Eyras, E. Leveraging transcript quantification for fast computation of alternative splicing profiles. *RNA* **21,** 1521–1531 (2015).

44. Trincado, J. L. *et al.* SUPPA2: Fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol.* **19,** (2018).

45. Li, Y. I., Sanchez-Pulido, L., Haerty, W. & Ponting, C. P. RBFOX and PTBP1 proteins regulate the alternative splicing of micro-exons in human brain transcripts. *Genome Res.* **25,** 1–13 (2015).

46. Charest, A. *et al.* Fusion of FIG to the receptor tyrosine kinase ROS in a glioblastoma with an interstitial del(6)(q21q21). *Genes. Chromosomes Cancer* **37,** 58–71 (2003).

47. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455,** 1061–8 (2008).

48. Pugh, T. J. *et al.* The genetic landscape of high-risk neuroblastoma. *Nat. Genet.* **45,** 279–84 (2013).

49. Carithers, L. J. *et al.* A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreserv. Biobank.* **13,** 311–9 (2015).

50. Sebestyén, E. *et al.* Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. *Genome Res.* **26,** (2016).

51. Wang, X. *et al.* Detection of Proteome Diversity Resulted from Alternative Splicing is Limited by Trypsin Cleavage Specificity. *Mol. Cell. Proteomics* **17,** 422–430 (2018).

52. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10 (2011).

53. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).

54. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* (2017). doi:10.1038/nmeth.4197

55. Clamp, M. *et al.* Ensembl 2002: Accommodating comparative genomics. *Nucleic Acids Res.* **31,** (2003).

56. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5,** 113 (2004).

57. Eyras, E., Caccamo, M., Curwen, V. & Clamp, M. ESTGenes: Alternative splicing from ESTs in Ensembl. *Genome Res.* **14,** (2004).