SQANTI: extensive characterization of long read transcript sequences for quality control in full-length transcriptome identification and quantification

Manuel Tardaguila[1*], Lorena de la Fuente[2*], Cristina Marti[2], Cécile Pereira[1], Hector del Risco[1], Marc Ferrell[1], Maravillas Mellado[3], Marissa Macchietto[4], Kenneth Verheggen[5,6], Mariola Edelmann[1], Iakes Ezkurdia[7], Jesus Vazquez[7], Michael Tress[8], Ali Mortazavi[4], Lennart Martens[5,6], Susana Rodriguez-Navarro[9], Victoria Moreno[3], Ana Conesa[§1,2]

[1]Department of Microbiology and Cell Science, Institute for Food and Agricultural Sciences, University of Florida, USA

[2]Genomics of Gene Expression Laboratory, Centro de Investigaciones Principe Felipe (CIPF), Valencia, Spain

[3]Neural Regeneration Laboratory, CIPF, Valencia, Spain

[4]Department of Developmental and Cell Biology, University of California, Irvine, CA, USA

[5]VIB-UGent Center for Medical Biotechnology, VIB, Ghent, Belgium

[6]Department of Biochemistry, Ghent University, Ghent, Belgium

[7]Centro Nacional de Investigaciones Cardiovasculares CNIC, Madrid, Spain

[8]Centro Nacional de Investigaciones Oncologicas CNIO, Madrid, Spain

[9]RNA transport and metabolism Laboratory, CIPF, Valencia, Spain

[*]Joint first authorship

[§] Corresponding author: aconesa@ufl.edu; aconesa@cipf.es

ABSTRACT

High-throughput sequencing of full-length transcripts using long reads has paved the way for the discovery of thousands of novel transcripts, even in very well annotated organisms as mice and humans. Nonetheless, there is a need for studies and tools that characterize these novel isoforms. Here we present SQANTI, an automated pipeline for the classification of long-read transcripts that computes over 30 descriptors, which can be used to assess the quality of the data and of the preprocessing pipelines. We applied SQANTI to a neuronal mouse transcriptome using PacBio long reads and illustrate how the tool is effective in readily describing the composition of and characterizing the full-length transcriptome. We perform extensive evaluation of ToFU PacBio transcripts by PCR to reveal that an important number of the novel transcripts are technical artifacts of the sequencing approach, and that SQANTI quality descriptors can be used to engineer a filtering strategy to remove them. By comparing our iso-transcriptome with public proteomics databases we find that alternative isoforms are elusive to proteogenomics detection and are abundant in major protein changes with respect to the principal isoform of their genes. A comparison of Iso-Seq over the classical RNA-seq approaches solely based on short-reads demonstrates that the PacBio transcriptome not only succeeds in capturing the most robustly expressed fraction of transcripts, but also avoids quantification errors

caused by unaccounted 3' end variability in the reference. SQANTI allows the user to maximize the analytical outcome of long read technologies by providing the tools to deliver quality-evaluated and curated full-length transcriptomes. SQANTI is available at https://bitbucket.org/ConesaLab/sqanti.

INTRODUCTION

Alternative Splicing (AS) and Alternative Polyadenylation (APA) are among the most fascinating and challenging aspects of eukaryotic transcriptomes. AS and APA are considered major mechanisms to generate transcriptome complexity and thus expand proteome diversity of higher organisms[1–3]. These post-transcriptional mechanisms have been reported to play critical roles in differentiation[4–7], speciation[3,8] and multiple human diseases such as cancer[9–11], diabetes[12,13] or neurological disorders[14–18], and therefore play a fundamental role in the establishment of organismal complexity[3,19,20]. The genome-wide analysis of AS has been done primarily using first exon microarrays and more recently short-read RNA-seq. These two methods are effective for the identification of AS events such as exon skipping or intron retention and have established the involvement of AS in many biological processes. However, both technologies have serious limitations for the reconstruction of the actual expressed transcripts, as short reads break the continuity of the transcript sequences and fail to resolve assembly ambiguities at complex loci[21,22]. This impairs any studies that would catalogue specific transcriptomes, investigate cis-acting mechanisms within transcripts, infer open reading frames or understand functional aspects of isoform diversity.

There has been increasing interest in the application of single-molecule sequencing to obtain full-length transcripts in human and plants using the long-reads from PacBio[22–25], as these eliminate the need of short-read assembly to define the transcriptome composition. The Iso-Seq PacBio pipeline for transcriptome sequencing consists of first obtaining an enrichment of full-length cDNA using the ClonTech SMARTer protocol, followed by formation of single molecule SMRTbells with specific PacBio linkers, which are subsequently sequenced. PacBio reads are typically longer than the full-length cDNA sequence meaning that each molecule can go through several passes of sequencing, with the consensus called a Read of Insert (RoI), which is the current standard PacBio output. RoIs where both cDNA primers and the poly(A) can be identified are called Full-length (FL) reads, while those that miss any of these tags are deemed non Full-length reads. PacBio sequencing suffers, however, from a relatively high raw error rate (around 15%) and a lower throughput compared to Illumina that could compromise detection of poorly expressed transcripts and the quantification of expression levels. There are several described methods for PacBio error correction and transcript identification. Au et al[26], proposed a hybrid sequencing approach, where PacBio RoIs are first corrected with the more accurate Illumina reads using the computationally intensive LSC algorithm[27] and transcripts are called by a combination of direct detection and prediction (IDP) with short reads, using the reference genome as template. The TAPIS pipeline, does not need Illumina, but performs several rounds of mapping and correction of RoIs on the reference genome, with apparently similar error correction efficiency as a short-read based method[28]. Finally, the ToFU PacBio pipeline[29], obtains auto-clusters of FL and nonFL RoIs and then computes a consensus transcript sequence where errors are significantly reduced. In all cases comparison to the reference gene models serves to call known and novel transcripts.

All PacBio transcriptome papers discover thousands of new transcripts, propose some kind of classification scheme by comparing to a reference annotation and find that the majority of novel transcripts appear in known genes[23,25,26,28,30]. However, details on the number, quality and characteristics of these new calls can vary greatly. Sequencing the transcriptome of hESCs by long reads followed by IDP analysis identified over 2,000 novel transcripts (~30%) and discovered new genes that were proven to be functional[26]. Tilgner et al. found using PacBio sequencing of the GM128787 cell line about 12,000 novel transcripts fully supported by previous splice site annotations or Illumina reads, but did not study novel junctions in detail[25]. For the sorghum transcriptome, 11,342 (40%) novel transcripts were found by PacBio from a total of nearly 1M reads using a filter on splice junction quality (SpliceGrapher[31]), and 6/6 random transcripts were confirmed by PCR. Finally, the maize multi-tissue transcriptome analysis identified over 111,151 transcripts from 3.7M RoIs, most of them novel and tissue-specific[30]. The authors found that between 10% and 20% of the PacBio junctions lacked coverage by Illumina reads and < 1% were non-canonical[30], but do not report on the number of affected transcripts or validate any. In all these cases, an in-depth characterization of the novel transcripts and junctions that would reveal potential biases and justify analysis choices was missing. We believe that such analysis is important as a great variety of FL and nonFL RoIs typically map at each genome locus and different processing pipelines can result in significantly different final transcript calls. As an example, sequencing the mouse neural transcriptome with PacBio, we obtained ~ 80,000, 12,000 and 16,000 different transcripts when applying Tapis, IDP or the ToFU pipeline, respectively. Implementing a comprehensive, quality aware analysis of PacBio reads is fundamental at a time when long read transcriptome sequencing is becoming more popular and important conclusions on transcriptome diversity will be drawn from these data.

In this work, we present SQANTI (Structural and Quality Annotation of Novel Transcript Isoforms), a pipeline for the analysis of long-read transcriptomics data that calculates up to 35 different descriptors of transcript quality and creates a wide range of summary graphs to aid in the interpretation of the sequencing output. We apply SQANTI to the analysis of the mouse neural transcriptome and illustrate how the tool is useful to characterize transcript types. We include thorough RT-PCR validation that verifies the expression of many novel transcripts but also reveals that an important fraction of the novel sequences are presumably bioinformatics or retrotranscription artifacts that can be removed by using SQANTI descriptors. Moreover, we confirmed the translation of a subset of alternative and novel transcripts by proteogenomics. Finally, we demonstrate that non-annotated variability at 3' ends of expressed transcripts might be confounding transcriptome quantification by short-reads alone. Our results indicate that using a full-length, experiment specific transcriptome as a reference solves this problem and improves accuracy of quantification estimates. Our work confirms the potential of long-read sequencing for precise characterization of the transcriptome complexity provided appropriate preprocessing steps are applied.

RESULTS

Experimental design and transcriptome sequencing

Full-length cDNA from Neural Progenitor Cells (NPCs) and Oligodendrocyte Precursors (OLPs), two biological replicates each, was obtained and split to prepare Illumina and PacBio sequencing libraries (Figure 1A). PacBio

sequencing was performed according to the Iso-Seq protocol to generate around 0.6 M RoIs per sample for a total of 2.2M RoIs. Illumina sequencing resulted in approximately 60 M reads per sample. All PacBio RoIs were joined and processed by the ToFU pipeline[29] to obtain a total of 16,104 primary PacBio transcripts. Alignment of the ToFU transcripts against the mouse reference genome (GMAP[32], assembly mm10) showed an average percentage of coverage and identity above 99.8%, indicating that most sequencing errors were corrected by the ToFU clustering approach. However, small indels (average size ~ 1.2 nts) were still detected in 56.2% of the transcripts. To tackle this problem we first attempted to correct indels with matching Illumina short reads using Proovreads[33]and LSC[27]. Although the number of transcripts with at least one indel decreased to 2,550, this was still unsatisfactory for ORF prediction. Instead, transcripts were corrected using the reference genome sequence (Figure 1A). By virtue of this strategy, all indels were removed and we obtained the *corrected PacBio transcriptome*.

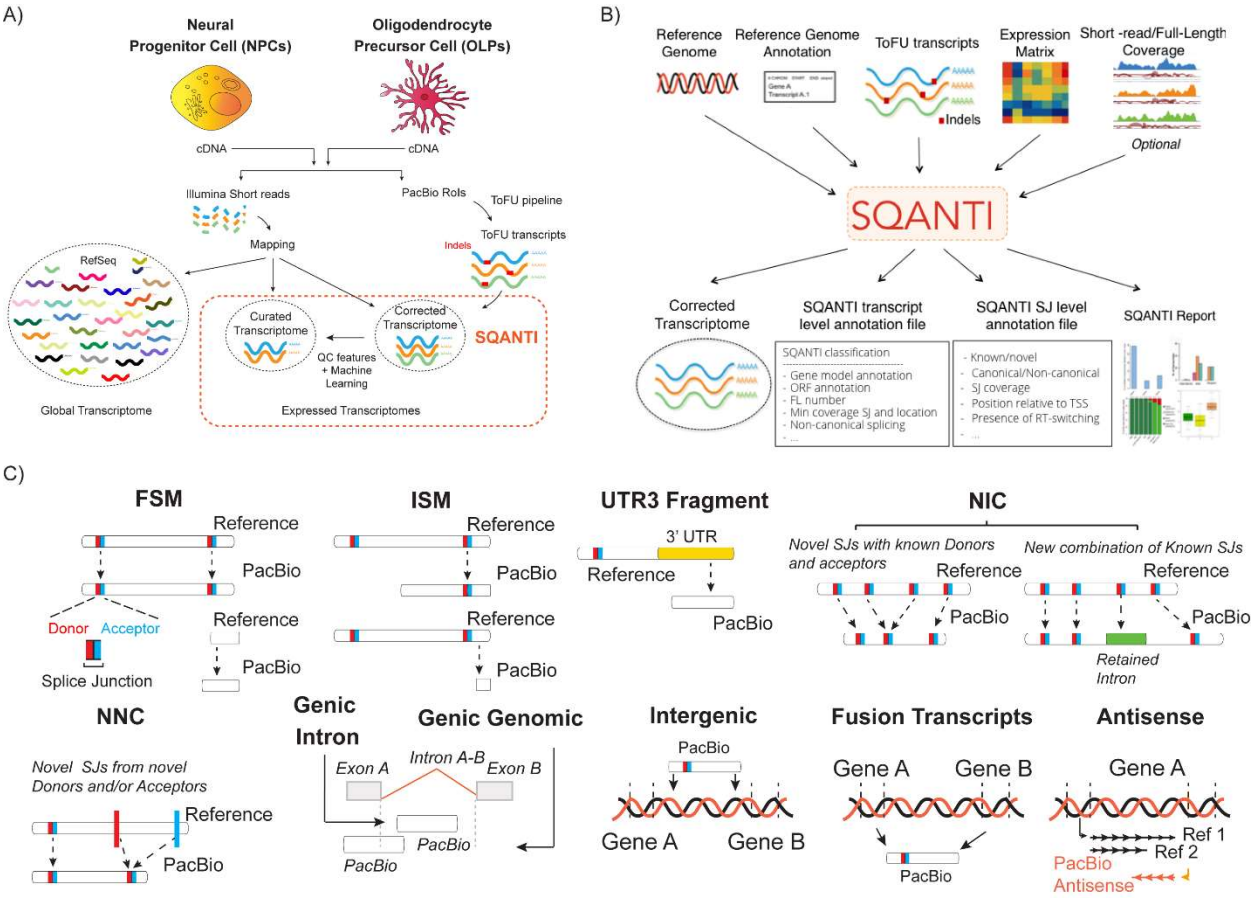

Figure 1. Overview of the experimental model and SQANTI analysis. A) Experimental system and data processing pipeline. RNA isolated from Neural Progenitor Cells (NPCs) and Oligodendrocyte Precursor Cells (OPCs) was retrotranscribed separately into cDNA, and sequenced both by long-read PacBio and short-read Illumina technologies. All PacBio RoIs were joined and processed by the ToFU pipeline to obtain consensus transcripts. Residual (indel) errors were removed by comparison to the reference genome to generate a corrected transcriptome and false transcripts were filtered out using a machine learning approach based on SQANTI QC features to result in a curated transcriptome. Illumina short reads were mapped against the RefSeq murine transcriptome annotation, the corrected and the curated PacBio transcriptomes. **B)** Workflow of SQANTI. SQANTI uses as input files: a fasta file with non-corrected transcript sequences, a transcript expression matrix, the fasta file with the reference genome and its gtf annotation file. Optionally the user can provide a text file with the number of Full-Length reads per transcript and the coverage at the splice junctions by short reads. SQANTI returns a reference-corrected transcriptome, a transcript-level and a

junction-level annotation files with structural and quality descriptors, and a pdf with graphs summarizing SQANTI data. **C)** SQANTI classification of transcripts according to their splice junctions and donor and acceptor sites. Splice donors and acceptors are indicated in red and blue respectively. SJ= splice junction, FSM=Full Splice Match, ISM=Incomplete Splice Match, NIC=Novel in Catalog, NNC=Novel Not in Catalog.

Transcript classification based on splice junctions

The SQANTI pipeline was developed for an in-depth characterization of PacBio transcripts. SQANTI takes as input information on transcript expression and reference annotations, and returns a reference-corrected transcriptome together with a wide set of transcript and junction descriptors which are further analyzed in several summary plots to assess the quality of the data (Figure 1B). Supplementary Tables 1 and 2 describe in detail the set of transcript descriptors computed by SQANTI.

A hallmark of the SQANTI analysis is the classification of transcripts based on the comparison of their splice junctions with the provided reference transcriptome to reveal the nature and magnitude of the novelty found by long-read sequencing. For our data we used as reference a non-redundant combination of the RefSeq and Ensembl mouse genome annotations, although other references may be provided by the user. PacBio transcripts matching a reference transcript at all splice junctions are labeled as Full Splice Match (FSM, Figure 1C), while transcripts matching consecutive, but not all, of the splice junctions of the reference transcripts are designated as Incomplete Splice Match (ISM, Figure 1C). Monoexonic transcripts matching a monoexonic reference were included in the FSM category whereas those matching a multiexonic reference were placed in the ISM group (Figure 1C). ISM transcripts contained by more than a 95% of their sequences within a reference 3'UTR are labeled as UTR3 Fragment (Figure 1C). Furthermore, SQANTI classifies novel transcripts of known genes into two categories: Novel in catalog (NIC) and Novel not in catalog (NNC, Figure 1C). NIC transcripts contain new combinations of already annotated splice junctions or novel splice junctions formed from already annotated donors and acceptors. NNC transcripts use novel donors and/or acceptors. Novel genes are classified as "Intergenic" transcripts, if lying outside the boundaries of an annotated gene, and as "Genic intron" transcripts if lying entirely within the boundaries of an annotated intron (Figure 1C). In addition, "Genic genomic" transcripts are monoexonic transcripts with partial exon and intron overlap in a known gene (Figure 1C). Finally, SQANTI labels Fusion transcripts (transcript spanning two annotated loci), and Antisense transcripts (transcripts being expressed from the contrary strand of the annotated coding transcripts in a given locus) (Figure 1C). Our corrected neural PacBio transcriptome contained a total of 16,104 transcripts resulting from the expression of 7,704 different genes. Following the SQANTI classification, transcripts mapping a known reference (FSM, ISM and UTR3 Fragment) accounted for 60% of the transcriptome, novel transcripts of known genes (NIC, NNC) made up 35.6% of our sequences while novel gene transcripts (Intergenic and Genic intron categories) represented about 2.3% of our data (Supplementary Figure 1A).

An important advantage of full-length transcript sequencing is that the prediction of ORFs, 5'UTR and 3'UTRs is greatly facilitated. SQANTI implements the GeneMarkS-T[13] (GMST) algorithm to predict ORFs from transcript sequences which showed highly reliable protein prediction in our data (Supplementary methods and Supplementary Figure 1B-D). GMST found 11,999 non-redundant ORFs within a total of 14,395 coding transcripts while 1,709 transcripts were predicted to be "ORF-less". The great majority of FSM, ISM, NIC and NNC transcripts were

predicted to have an ORFs (97%, 90%, 87.8% and 92.8%, respectively), while the rest of transcript categories, including UTR3 Fragments, were mostly ORF-less.

Descriptive analysis of transcriptome complexity and full-lengthness made easy by SQANTI

A fundamental goal of long-read transcriptome sequencing is to capture the extent of transcriptome complexity and to obtain full-length transcripts. SQANTI includes all basic graphics to readily study these aspects. SQANTI calculates transcript length distribution, reference transcript length, number of supporting FL reads, transcript expression, reference coverage at both 3' and 5' ends, number of exons and number of transcripts per gene (Supplementary Table 1). Moreover, analyses are provided with the transcript classification breakdown, which adds an extra layer of understanding on the quality of the sequencing results. For example, we hypothesize that ISM transcripts are a combination of potentially real shorter versions of long reference transcripts as well as partial fragments resulting from incomplete retrotranscription or mRNA decay, while UTR3 Fragment transcripts might mostly be composed of the second. Indeed, SQANTI analysis shows that PacBio transcripts classified as ISM matched reference transcripts that were longer (Figure 2A) and had more exons (Supplementary Figure 2A) than FSM sequences. Moreover, UTR3 Fragment transcripts matched the longest reference transcripts (Figure 2A) suggesting their enrichment in retrotranscription fragments. Applied to our data we observed that, although all groups showed similar distribution of transcript lengths (Figure 2B), UTR3 Fragment, Genic Genomic, Intergenic and Genic Intron transcripts were almost entirely composed by monoexon transcripts, which was not true for other transcript categories (Supplementary Figure 2B). Remarkably, only 13.8% of novel genes had splice junctions and most of them (98.2%) expressed just one transcript (Figure 2C and Supplementary Figure 2C). In addition, SQANTI calculates the extent of overlap between sequenced and reference transcript at 3' and 5' ends as a proxy to evaluate transcript full-lengthness. This analysis makes sense for FSM transcripts, for which a reference with an identical splice pattern exists. In order to exclude 3'/5' overlap differences due to alternative polyadenylation or alternative use of TSS we restricted our analysis to matches within the 100 most extreme nucleotides of the reference. As expected, the majority of our FSM transcripts showed a complete or close to complete 3' end overlap with the 3' end of the matched reference transcript: 76% had an exact 3' end match and 16% were within 20 nts upstream of the annotated 3' end (Figure 2E). This contrasts with the 35% of FSM transcripts showing a complete overlap with their reference 5' ends and the 50% falling short by 40 to 100 nts (Figure 2F). This result is in agreement with the used cDNA library preparation strategy and ToFU analysis parameters that require identification of poly(A) tails to call FL reads, but have less control over completeness at 5' ends. Interestingly, 851 and 1,361 FSM transcripts had 3'end and 5'end positions further down/upstream of the matched reference transcript, while 1,610 and 1,439 of our FSM sequences, lacked 3' and 5' overlap, respectively, of more than 100 nts. These cases might represent alternative polyadenylation/alternative TSS events. Finally, SQANTI descriptive graphs reveal differences between transcript categories at expression features. For example, transcript expression level and number of supporting FL reads tend to decrease from FSM to NIC and NNC transcripts (Figure 2D and Supplementary Figure 2D) and are low for novel genes compared to annotated genes (Supplementary Figure 2E and 2F), which shows that novel transcripts have generally lower expression levels than those already identified in reference databases.

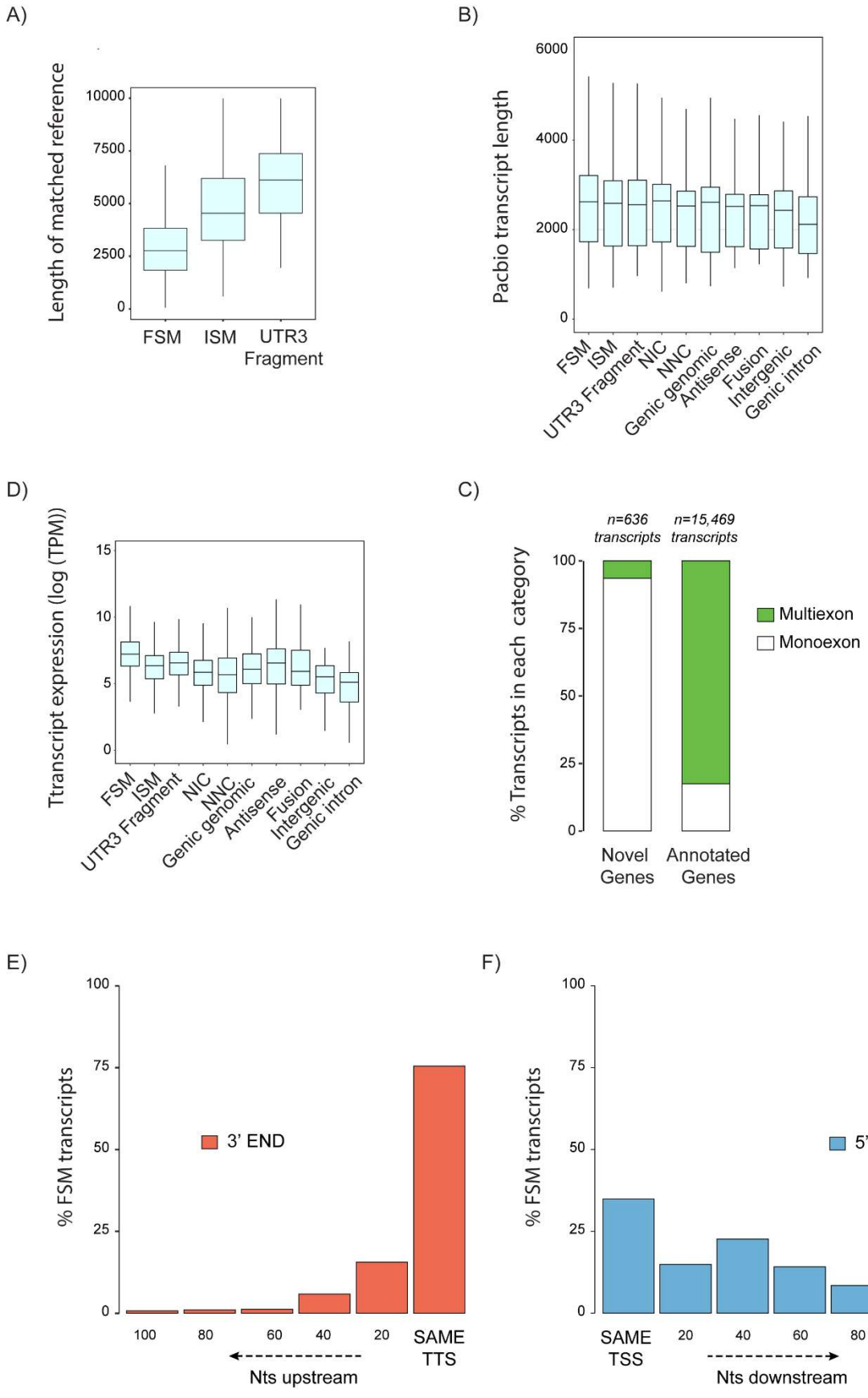

Figure 2. SQANTI characterization of the corrected PacBio transcriptome. A) Length of the reference transcripts to which FSM, ISM and UTR3 Fragment PacBio transcripts match. **B)** Length of PacBio transcripts by SQANTI categories. **C)** Percentage of monoexonic and multiexonic transcripts for transcripts belonging to novel genes and annotated genes. **D)** Transcript expression across SQANTI categories. **E)** Overlap at 3' and **F)** 5' ends between the FSM transcripts and their respective matched reference transcripts. TTS = Transcription Termination Site, TSS = Transcription Start Site.

In summary, the descriptive analysis framework provided by SQANTI readily indicates that our neural mouse transcriptome, obtained by PacBio single molecule sequencing, is effective in recovering full-length transcripts and shows an important level of novelty (~ 40%) with respect to the reference mouse transcriptome both because of novel splicing events and 3'/5' end length variation. Transcript diversity is more important than the presence of novel genes, which represents only a small fraction of the expressed mRNAs. However, novel transcripts tend to be less expressed than annotated transcripts indicating that, generally, less novelty is to be expected for major transcripts.

Evaluation of transcripts according to their splice junctions

In order to better understand the sources of novel transcripts SQANTI further analyzes their splice junctions. Splice junctions can be divided into canonical and non-canonical according to the two pairs of dinucleotides present at the beginning and at the end of the introns encompassed by the junctions. The combination of GT at the beginning and AG at the end of the intron is found in 98.9% of all the introns in the human genome[35]. We considered GT-AG as well as GC-AG and AT-AC as canonical splicing (altogether found in more than 99.9% of all the human intron [35,36]), and all the other possible combinations as non-canonical splicing. SQANTI also allows users to provide their own set of canonical junctions. At the same time, SQANTI subdivides splice junctions between known, if they were present in the reference, and novel, if they were not.

In our mouse neural data, the ratio of canonical versus non-canonical splicing events fitted the expected genome proportions: out of 141,332 known splice junctions, 99.9% were canonical and 0.1% (185) were non-canonical. However, novel splice junctions showed a much different distribution: out of 3,837 novel splice junctions, 69% were canonical and 31% (1,188) were non-canonical. When analyzed across the different SQANTI categories, non-canonical splicing was maintained at low rates in FSM (0.1%) and ISM (0.25%) transcripts which was expected as both are formed purely by known splicing events (the few UTR3 Fragment transcripts having splice junctions showed a 100% of canonical splicing) (Figure 3A). In NIC transcripts, where novel combinations of known splice junctions or novel splice junctions deriving from annotated donors or acceptors are present, the percentage of non-canonical splicing was 0.15%. In all cases, these non-canonical junctions were already known in the reference, meaning that all novel junctions in this category were canonical (Figure 3A). However, in NNC transcripts, characterized by the introduction of alternative donors and/or acceptors, we found 1,155 novel non-canonical junctions, which represented 4.45%, of the splice sites in these transcripts (Figure 3A). Moreover; Intergenic, Genic Intron, Genic Genomic and Antisense transcripts, despite rarely being multiexonic, showed relatively high percentages of non-canonical splice junctions with 2.32%, 7.28%, 21.57% and 32.65% respectively (Figure 3A). This unusual high level of non-canonical junctions suggests that experimental artifacts might be accumulating in

these categories. Furthermore, when the percentage of transcripts showing at least one non-canonical splice junction was considered, the proportion of NNC affected compared to NIC transcripts became more evident, 41.5% vs 1.47%, respectively, arguing that this category of transcripts needed deeper inspection.

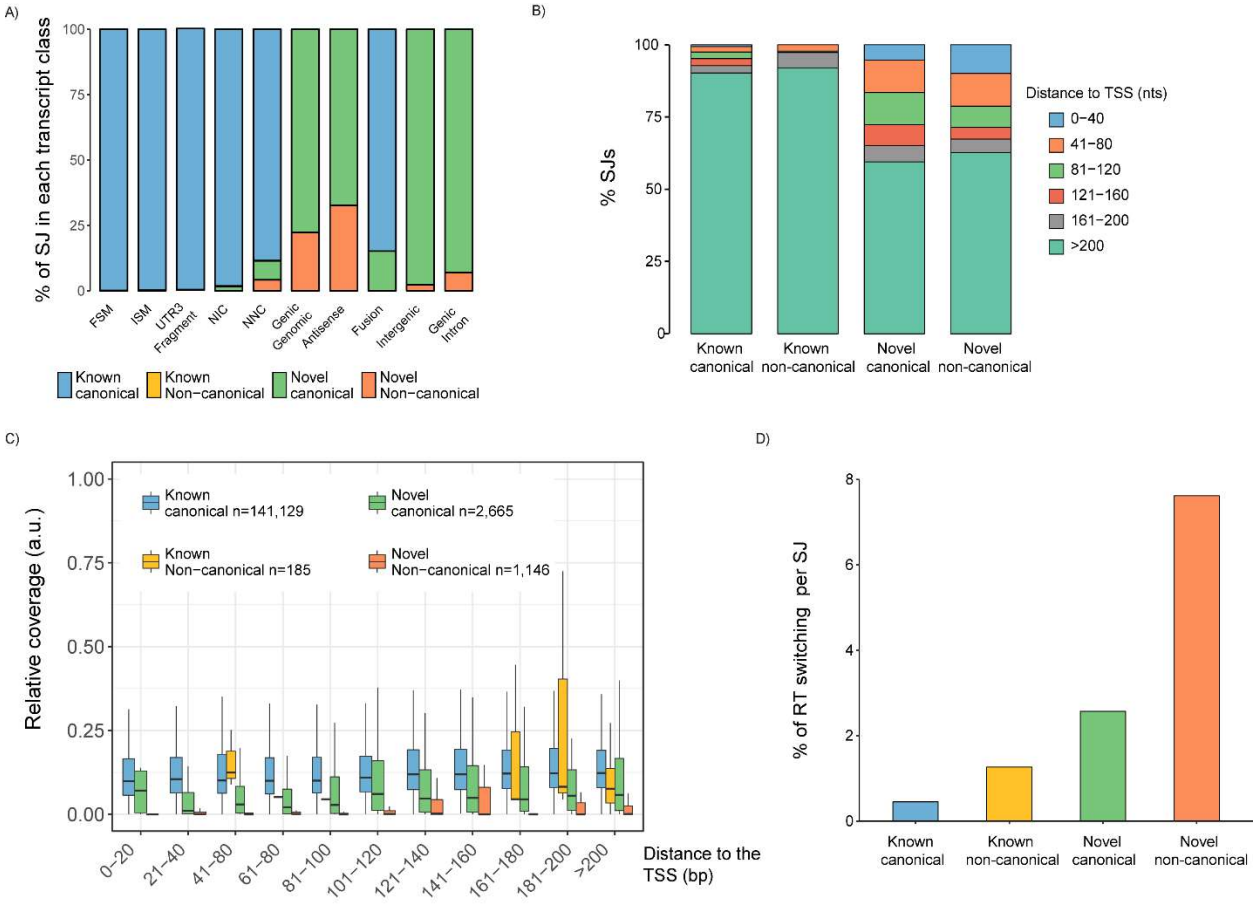

Figure 3. Splice junction's characterization in the corrected PacBio transcriptome. A) Distribution of Splice Junction (SJ) types across SQANTI categories. NNC, Genic genomic, Antisense, Intergenic and Genic intron are enriched in non-canonical SJs. n= 76,757 SJ for FSM, n= 13,784 for ISM, n=18 for UTR3 Fragment, n=27,368 for NIC, n=26,509 for NNC, n=51 for Genic Genomic, n=49 for Antisense, n= 494 for Fusion, n=86 for Intergenic and n=55 for Genic Intron. **B)** Distribution of the SJs according to their distance to the Transcription Start Site (TSS). **C)** Relative coverage by short-reads of SJs as a function of their class and distance to the TSS. a.u. = arbitrary units. **D)** Detection of RT switching direct repetitions by SQANTI algorithm across SJ types.

Next we investigated the position of novel junctions with respect to transcript 5' ends. Interestingly we found, that although novel junctions could appear at any position in novel transcripts, there was a higher concentration of occurrences towards 5' ends which is not observed for known - whether canonical or not - junctions (Figure 3B, Fisher's exact test $p < 2.2 \times 10^{-16}$). This could either be the consequence of unannotated variability at 5' ends or higher accumulation of errors due to lower sequence support. The ToFU pipeline is more permissive with clustering conditions at transcript ends (E. Tseng, personal communication), which accounts for a higher probability of errors at these areas. Coverage by Illumina has been used to support novel junctions called by PacBio[26]. However, Illumina

reads are not always equally distributed along the transcript length and are often less abundant towards 5' ends, providing less support for junction validation. SQANTI examines the level of coverage by Illumina reads of known and novel junctions as a function of their distance to the 5' end of the PacBio transcript. We found that, as suspected, splice junction support by short reads decreased towards the 5' end of the transcripts, but was significantly higher for known junctions (Figure 3C, Wilcoxon test $p < 2.2 \times 10^{-16}$). Novel canonical junctions were in general less supported but were still significantly more supported than novel non-canonical junctions, which had hardly any supporting reads if located within the first 120 nts of the transcript 5' end (Figure 3C, Wilcoxon test $p < 2.2 \times 10^{-16}$).

Another possible explanation for the high rate of non-canonical splicing in the NNC group is Reverse Transcriptase template switching (RT switching). RT switching is an experimental artifact resulting from secondary structures in the RNA template that makes continuous portions of the template inaccessible to the retrotranscriptase enzyme. These gaps originate at the cDNA synthesis step and are subsequently interpreted as splicing events, which, due to their non-splicing origin, are enriched for non-canonical junctions[36]. A hallmark of RT switching is the presence of a direct repeat between the upstream mRNA boundary of the non-canonical intron and the intron region adjacent to the downstream exon boundary[36]. SQANTI incorporates an algorithm to locate these direct repeats, which confirmed the enrichment of RT switching detection in novel non-canonical splice junctions (Figure 3D, Fisher's exact test $p< 2.2 \times 10^{-16}$) and in NNC compared to NIC transcripts (7.24% versus 1.98%). Described RT switching events affect minor isoforms of genes co-expressed with a major isoform that serves as the template for the intramolecular switching[36]. Accordingly, we found that NNC transcripts are enriched for being minor transcripts of highly expressed genes (Supplementary Figure 2G and 2H). Moreover, RT switching direct repeats were enriched in the NNC minor transcript compared to the major FSM transcript of the gene (Supplementary Figure 2I).

Altogether, the SQANTI framework analyses suggests that at least a fraction of the novel transcripts found by PacBio sequencing could originate by technical artifacts at the step of cDNA library construction or by less confident sequencing data at the 5' ends of transcripts.

PCR validation of PacBio transcripts

To shed light into the correct detection of transcripts by PacBio sequencing we performed RT-PCR amplification for a total of 57 mRNAs encompassing different SQANTI categories: 24 FSM (3 with non-canonical splice sites, Supplementary Figure 3A1-4), 9 NIC (Supplementary Figure 3B1-2), 21 NNC canonical (14 of them containing at least one non-canonical splice junction, Supplementary Figure 3C1-4) and 3 fusion transcripts (Supplementary Figure 3D). Importantly, we performed RT-PCRs both on the ClonTech oligo(dT) enriched full-length cDNAs used for PacBio sequencing and on new cDNA retrotranscribed at 42 ºC and 50 ºC using random hexamers rather than oligo(dT). The rationale behind this approach was to test whether novel transcripts could have been spuriously generated by RT switching-like mechanisms at the retrotranscription step of the PacBio protocol. Since higher temperature and/or the use of random hexamers would complicate the formation of secondary structures in the RNA template, retrotranscription artifacts would be less favored in these conditions.

We validated by RT-PCR 23/24 of the FSM, including the 3 cases with non-canonical junctions, (Figure 4A1) and 7/9 of the NIC transcripts, indicating very high identification rate by the PacBio platform within these categories. However, the NNC category gave very different results: 17/21 NNC transcripts were not confirmed by RT-PCR (Figure 4A2). Interestingly, half of the PCR confirmed NNC transcripts (2/4) did contain non-canonical junctions (Figure 4A3). Moreover, there were 6 NNC cases in which amplification with oligo(dT) cDNA resulted in a band of the expected size, while the band was no longer detectable when random hexamers were used to obtain cDNAs for PCR (Figure 4A4). Such loss of detection upon change of PCR conditions was never observed for confirmed transcripts in any of the other categories, which remained consistently positive in all PCR tests. In addition, we re-assayed with random hexamers some of the transcripts that were not amplified in the oligo(dT) RT and we did not observe rescue of amplification for any (Supplementary Figure 3C). Lastly, we confirmed 1/3 of the Fusion transcripts. Table 1 summarizes the results of the PCR validation experiment. Details can be found in Supplementary Table 2.

	oligo(dT)			**Random Hexamers**		
Transcript Type	**Positive**	**Negative**	**Total**	**Positive**	**Negative**	**Total**
FSM	23 (3 nc)	1	24	5 (3 nc)	0	5
NIC	7	2	9	7	0	7
NNC	10	11	21	4 (2 nc)	6	10
Fusion	1	2	3	1	0	1

Table 1. Summary RT-PCR validation. nc: transcript with non-canonical junctions.

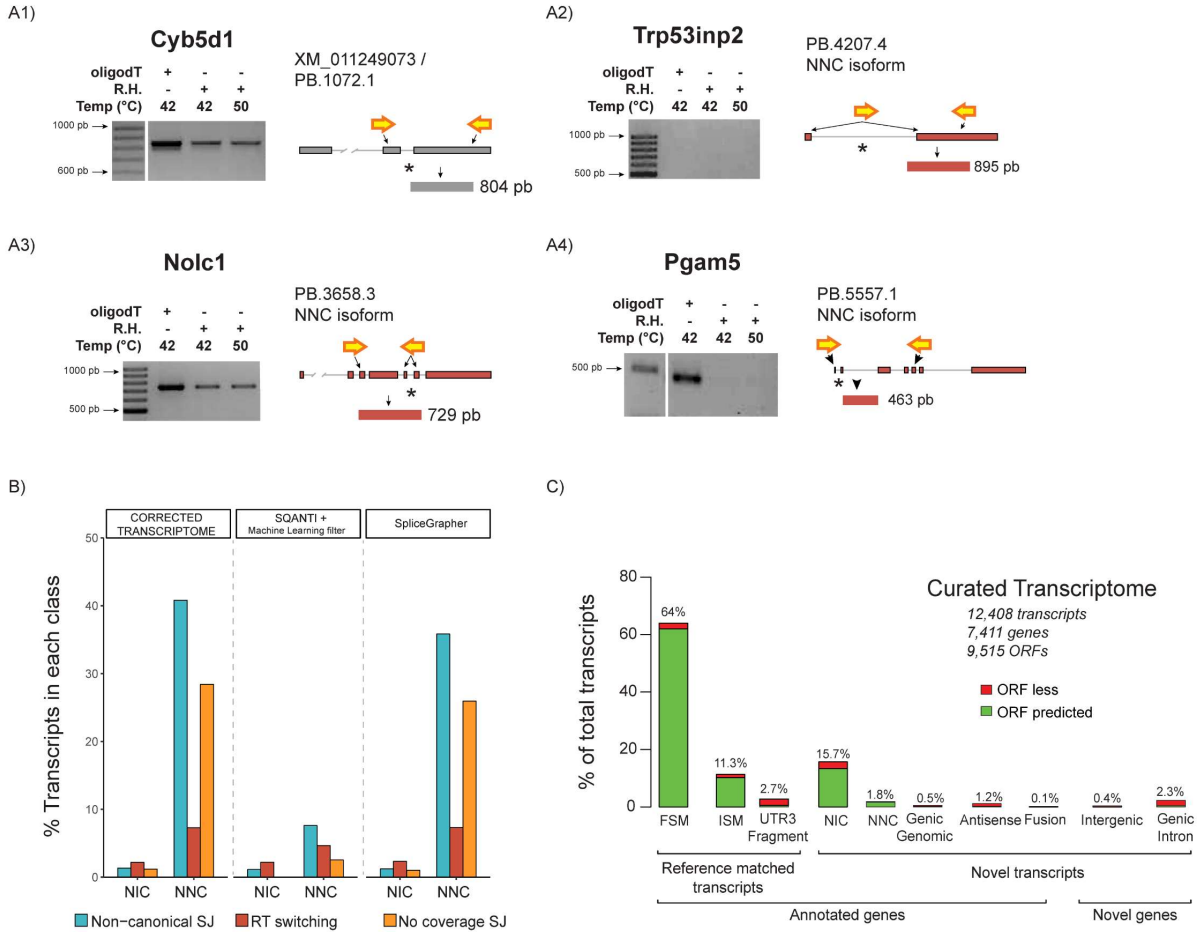

Figure 4. RT-PCR validation of PacBio transcripts. A) Representative examples of RT-PCR validation experiments. Three PCR conditions were assessed: oligo(dT) template at 42 ºC and Random hexamers (R.H.) template at 42 ºC and at 50 ºC. **A1)** Example of a FSM transcript with a non-canonical SJ successfully amplified at each PCR condition, **A2)** Example of a NNC transcript with a non-canonical SJ that failed to be amplified, **A3)** Example of a NNC transcript with non-canonical SJ successfully amplified at each PCR condition **A4)** Example of NNC transcript with non-canonical SJ amplified at oligo(dT) but not at RH conditions **B)** Distribution of three *bad features of splicing* across NIC and NNC transcripts in the corrected transcriptome, the SQANTI curated transcriptome and after filtering by SpliceGrapher. **C)** Transcript class and ORF composition of the curated PacBio transcriptome. FSM matches are depicted in grey and novel isoforms in red.

Using SQANTI features to build a quality control filter for PacBio artifacts

The previous analyses established that an important fraction of novel PacBio transcripts could not be validated by PCR. Moreover, the SQANTI QC analysis revealed distinct signs of low quality within the NNC transcript category, such as non-canonical sites, RT switching direct repeats or low expression. However, none of the features considered by SQANTI were a clear hallmark of these novel junctions. For example, a few known junctions lack Illumina coverage (61 out of 141,334), while many of the novel non-canonical junctions did have supporting Illumina reads (655 out of 1,188) and two of them were confirmed by RT-PCR. RT switching direct repeats and low expression values accumulated in transcripts not detected by PCR, but were not exclusive to them. We used the

quality features evaluated by SQANTI and the results of our RT-PCR validation experiments to develop a machine learning-based quality filtering strategy for PacBio novel calls. Considered descriptors are listed in Supplementary Table 1 and comprise expression related values, structural features and features related to the quality of the alignment to the reference transcriptome. We defined as true-positives those transcripts verified by PCR (n=35), true negatives were the transcripts with negative RT-PCR amplification (n=22), applied Random Forest[37] as classification method and evaluated performance by cross-validation (Methods). The resulting classifier contained 8 selected descriptors and was successful in selecting true positive transcripts with an accuracy of 0.813 (AUC=0.902, specificity=0.813, sensibility=0.830), which can be considered a very good performance (Supplementary Figure 4A and 4B). The features selected by the classifier are shown in order of importance in Supplementary figure 4C. Interestingly, five out of eight discriminative variables selected by the filtering were associated with transcript expression, namely number of FL reads, lowest Illumina coverage at junction, minimum sample coverage, minimum coverage of an exon and transcript expression value, suggesting that expression patterns are within the most definitive characteristics to call *bona fide* novel transcripts. Moreover, position data was also relevant for the classification: coverage of the first and last exons were selected variables. The last classifier feature, Bite, indicates the transcript skips consecutive exons and has donor/acceptor sites inside a known exon, which we interpret as an indication of the novel splice junctions could have been spuriously formed from secondary RNA structures. Applied to our data, this method discarded 3,696 transcripts (2,467 NNC, 1,095 NIC) as being artifacts and retained as total number of 2,397 new transcripts. Importantly, after filtering, the percentage of non-canonical junctions within NNC transcripts was brought down from 41% to 7%, NNC transcripts with no Illumina junction support were lowered to 2% and the fraction of NNC with signs of RT switching was lowered to 4.4% (Figure 4B). We compared this result with SpliceGrapher[31], an algorithm used by the TAPIS pipeline which filters splice junctions from gapped alignments to remove artifact reads. SpliceGrapher filtered out 1,010 transcripts (716 NNC transcripts and 197 NIC) without affecting significantly the incidence of any of these low quality features in the NNC category (Figure 4B). Furthermore, SpliceGrapher accepted 17 out of the 22 transcripts that were negative by RT-PCR. We concluded that SQANTI's features provided an efficient framework to remove transcripts likely to be sequencing artifacts. The SQANTI development implements a function that uses this set of features to train and filter any user-provided transcriptomics dataset analyzed by SQANTI.

The adjusted percentages of each transcript category in our *curated transcriptome* were: 64% FSM, 11.3% ISM, 2.7% Fragment UTR3, 15.7 % NIC, 1.8% NNC, 0.5% Genic genomic, 1.2% Antisense, 0.1% Fusion transcripts, 0.4% Intergenic and 2.3% Genic Intron (Figure 4C). In our curated transcriptome 9,680 transcripts (78%) are in the known category, 2,397 (19.3%) are novel transcripts and 331 (2.7%) fall within novel genes. These transcripts were the product of 7,411 genes and resulted in 9,815 different ORFs.

Analysis of peptide support for the curated transcriptome.

Most of the novel transcripts were predicted to have ORFs that contained novel amino acid stretches when compared to principal isoforms. We sought to investigate whether peptide data present in public proteomics databases could support these findings. In order to do this we first created a non-redundant ORF database of public mouse proteins

and proteins in our neural data, and classified each protein as Principal Isoform ORF (PI-ORFs, n=4,581) if annotated as such in the Principal Isoform predictor APPRIS[34], Alternative ORF (Alt-ORF, n=2,127), if present in Ensembl or RefSeq but not being PI, and Novel ORF (Novel-ORF, n=1,472), if the protein would be coded by NIC or NNC transcripts present only in our mouse PacBio data. For each predicted protein, we performed an *in-silico* trypsin digestion and selected unique peptides that would unequivocally identify each ORF. We analyzed theoretical peptides for those genes identified in our mouse transcriptome that had more than one isoform annotated in Ensembl (v80). The percentage of ORFs predicted to be identifiable by unique peptides was highest for the PI-ORFs (73.7% or 2,591), followed by the Novel-ORFs (39.8% or 586) and was lowest for Alt-ORFs (30%, or 642). The majority of Novel-ORFs and Alt-ORFs were predicted to have only one unique peptide, while this was only the case for 14.3% of the PI-ORFs (Supplementary Figure 5A). Conversely, most PI-ORFs were predicted to contain 6 or more discriminating peptides and this was true for only 6.9% of Alt-ORFs and 10.2% of Novel-ORFs (Supplementary Figure 5A). This higher rate of unique peptides in PI-ORFs was expected as the mouse genome contains a significant number of genes in which alternative isoforms have only partial sequences and the APPRIS PI is often the longest ORF in a gene. Consequently, proteins deemed as PI are expected to be easier to detect by protein digestion approaches than alternative isoforms.

We then screened public databases for the presence of unique peptides associated to our set of ORFs. Two separate approaches were conducted: a *Neural tissue* approach, comprising one proteomics study of mouse neural tissue and another study of the mouse neural secretome, and an *All tissue* approach comprising peptides from 36 proteomics studies carried out on a variety of murine tissues but excluding the two ones used in the first approach. Overall, we detected at least one unique peptide for 77.9% of the PI-ORFs predicted to be identifiable, while this percentage went down to 20.56% and 8% for Alt-ORFs and Novel-ORFs, respectively. Most Alt- and Novel-ORFs had single unique peptide matches, while most PI-ORFs were found with multiple peptides (Supplementary Figure 5B). In part this is to be expected; the success of detection was significantly lower when the ORF was predicted to have only one unique theoretical peptide, and this was the case for the majority of Alt-ORFs and Novel-ORFs (Supplementary Figure 5C). At the same it is interesting to note that the agreement between the two proteomics screening approaches was much stronger for those proteins detected with two or more peptides (Supplementary Figure 5B). When ORFs were identified by a single peptide, the peptide was almost always present in just one of the two studies. Note that ORF detection by single peptide matches, similarly to transcript detection by single read counts, fall into the area of unreliable protein identification and therefore false discovery in these cases is not controlled[38]. This result confirms that the lower number of discriminating peptides in Alt and Novel ORFs versus their PI ORF counterparts impairs their detection by proteogenomics approaches, but other factors are also contributing, as Alt/Novel ORFs had lower unique peptide detection rates across all unique peptide ranges (Supplementary Figure 5C).

To understand if expression levels were playing a role, we evaluated the number of studies (PSM counts) supporting each ORF to find that on average Alt- and Novel-ORFs had 5-6 supporting studies (median=2) per detected unique peptide, while this number was nearly 10 for PI-ORFs (median=4.5), which is in agreement with the notion that PI-ORFs are ubiquitously expressed across tissues[34]. Interestingly, we found that PI-ORFs detected by unique peptides

in less than 5 proteomics studies had a significantly lower expression in our system than those found in more than 10 projects, and had similar expression levels as the transcripts coding for Alt- and Novel-ORFs (Supplementary Figure 5D). Altogether, these results indicate that direct detection in public proteomics databases of predicted coding products of novel and alternative transcripts is hampered by their lower expression pattern and an overall lower identifiability by unique peptides.

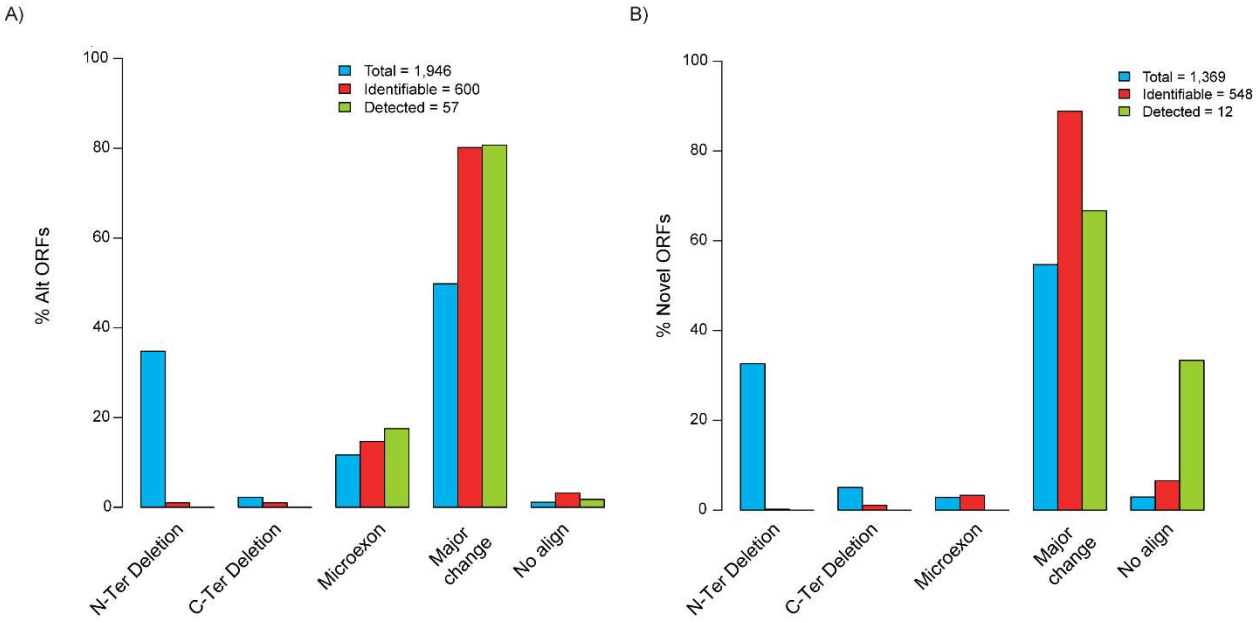

Figure 5. **Comparative analysis of protein differences between PI and non PI-ORFs.** Blue: ORFs computationally predicted in the curated transcriptome; red: ORFs predicted to be identifiable by unique peptides; green: ORFs detected in proteomics databases. **A)** Analysis for Alternative ORFs **B)** Analysis for Novel ORFs. ORFs positively identified by proteogenomics analysis were those with at least two Peptide Spectrum Matches (PSMs).

Finally, we evaluated the types of protein differences between alternative and principal isoforms for which peptide support was conclusive (minimum of 2 PSM counts per ORF, n=57 Alt-ORFs and n=12 Novel-ORFs), and compared them to the composition of our predicted transcriptomes. While our set of curated transcripts predicted that most alternative and novel ORFs distributed between N-terminal truncations, microexons (indels/substitutions up to 9 amino-acids, aas) and major changes (indels/substitutions of more than 9 aas with or without N-Ter/C-Ter truncations), the proteogenomics analysis failed to reveal most of these N-terminal differences and mostly found the major changes both for Alt- and Novel-ORFs (Figure 5), which is in agreement with a detection approach that relies on positive detection of unique peptides. Microexons were found mostly in Alt-ORFs (Figure 5A) while Novel-ORFs with no overlap to their PIs were found in the proteomics databases more than expected (Figure 5B), however this finding is supported by just a few ORFs and hence cannot be conclusive. Interestingly, although there was more than a 10-fold difference between the number of identifiable ORFs and those consistently identified in our proteomics screenings, there was a general agreement between the relative abundance of each type of protein differences among the two ORF sets, which suggests that the ORFs confidently identified by unique peptide matches could represent the actual diversity range of the alternative proteome.

Is long read sequencing improving transcriptome quantification?

While PacBio is effective in obtaining full-length transcripts and discovering novel transcript variants, the lower throughput of this technology is a problem for the detection of poorly expressed transcripts. Furthermore, several studies have shown that using a reduced, expressed transcriptome for short read mapping rather than the full transcriptome annotation improves quantification accuracy[26,39]. In addition, it is not yet fully clear if typical PacBio sequencing depth is sufficient to quantify gene expression or whether high-throughput Illumina is also required to obtain accurate transcript level estimates. In this section we analyze the effectiveness of PacBio sequencing regarding quantification of transcripts and compare to the typical RNA-seq approach based on Illumina short-read.

In order to gain insight into how transcriptome coverage between PacBio and Illumina differs, we mapped our short-reads to the mouse genome, the complete RefSeq transcriptome (~160,000 transcripts, from now onwards Global Reference, GlR) and our PacBio curated transcriptome (12,408 transcripts, from now onwards Expressed Reference, ExR). On average, 87% of our Illumina reads mapped to the mouse genome while 81.7% and 70.7% had a hit to the GlR and to our ExR respectively, indicating that only an 11% in transcriptional signal was missed when considering the ExR alone (Figure 6A.1). This difference in the number of mapped reads translated into a much bigger difference in the number of detected transcripts (30,071 versus 11,921 transcripts at a 1 count threshold), suggesting that GlR exclusive transcripts had little expression, which we confirmed after analyzing transcript expression levels (Figure 6A.2). At the gene level, GlR based quantification totally overlapped ExR except for 357 genes that were a combination of novel, fusion and other reference genes (Supplementary Figure 6A). A total of 3,447 transcripts that were absent from the GlR mapping were found using the ExR reference (~30% of total transcripts in ExR). Interestingly, 20.5% of these transcripts were annotated by Ensembl (Supplementary Figure 6B) while 7% were RefSeq transcripts. The rest of the ExR exclusive transcripts were novel transcripts (n=2,728), most of them NIC transcripts generated by new combinations of already known splice junctions (Supplementary Figure 6C). In addition, imposing a filter of 10 counts, eliminated most of the GlR-exclusive transcripts and made the number of transcripts and genes detected by the two mapping approaches similar (Supplementary Figure 6D). Note that a minimum of 10 counts is required by popular differential expression algorithms such as edgeR[40] to remove transcriptional noise. Furthermore, the proportion of genes with multiple transcripts was almost identical for the ExR and the GlR at this 10 count threshold (Supplementary Figure 6E). We concluded that, at reasonable sequencing depths for long and short reads technologies (0.6M and 60M, respectively), the PacBio transcriptome still captures nearly 90% of the transcriptional signal that Illumina would find and dramatically reduces calls of transcripts with very little expression that are at the margins of accurate quantification, while still rescuing transcriptional diversity not yet annotated by the reference databases.

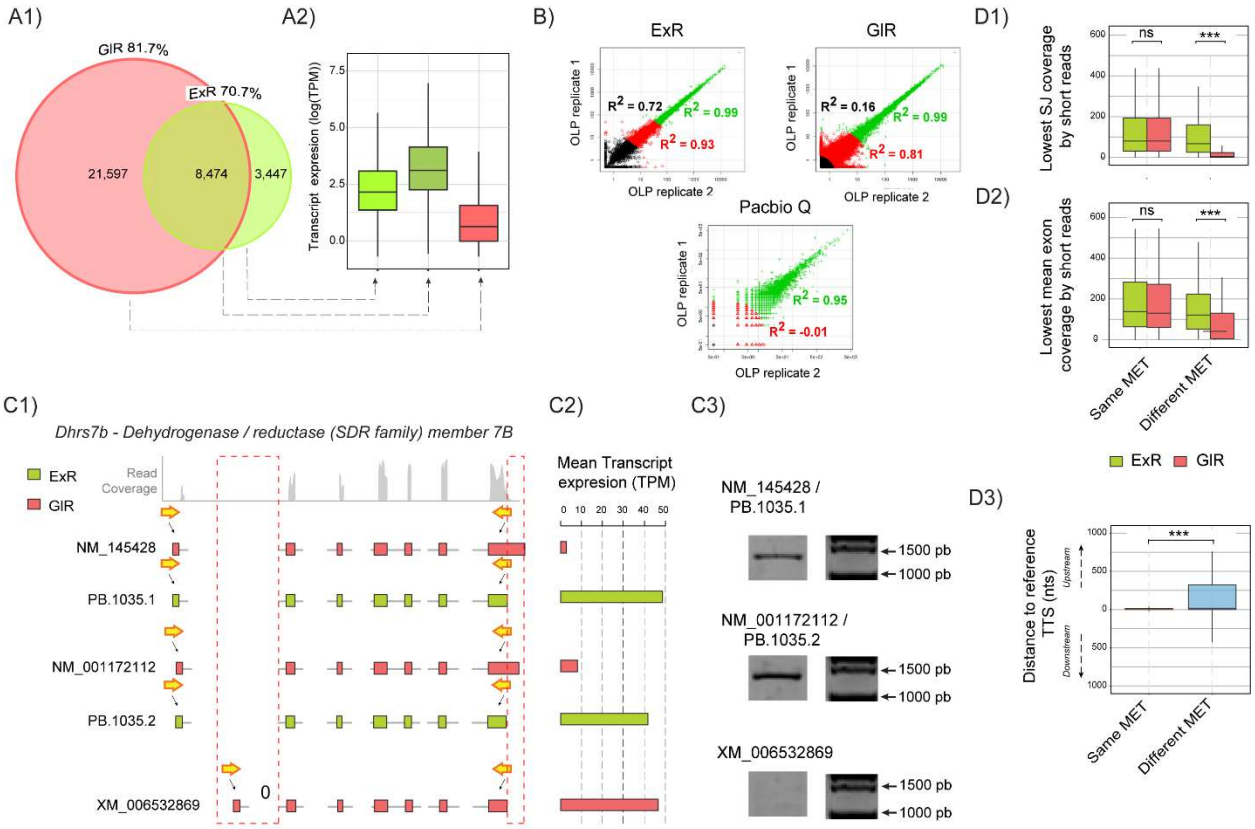

Figure 6. Short read mapping and transcript quantification using the Global Reference (GlR) and the Expressed Reference (ExR). A1) Venn diagram of the number of transcripts by short read mapping against GlR (red) and against ExR (green). Both quantifications overlapped in 8,474 transcripts (shaded area). The percentage of short reads mapped to each reference is indicated. **A2)** Boxplot of expression values of transcripts detected exclusively by GlR, ExR and by both quantification strategies. **B)** Correlation analysis between transcript expression values in OPC replicate samples when quantified by short-read ExR and GlR mapping, and by PacBio FL number. Correlation were computed at three bins of expression: low (black), medium (red) and high (green). **C)** Example of quantification in the GlR due to 3' end variability. **C1)** Transcripts associated to gene Dhrs7b according to PacBio sequencing (green) and by RSEM quantification using RefSeq (orange). The profile of mapping short reads at the DHrs7b locus is shown in grey. Red dashed boxes delineate exons without read coverage and 0 indicates splice junctions lacking any short read support. Position of transcript-specific primers are indicated by yellow arrows. **C2)** Average transcript expression levels according to ExR (green) and GlR (orange) quantification strategies. **C3)** PCR of DHrs7b transcripts with specific primers: PB.1035.1/ NM_145428 and PB.1035.2/NM_001172112 but not XM_006532869 were amplified. **D)** Analysis of the Most Expressed Transcript (MET) in genes with MET differences between ExR and GlR quantifications. Significant differences were calculated by Kruskal-Wallis test and $p < 0.001$. **D1)** Lowest SJ coverage by short-reads in METs. **D2)** Lowest mean exon coverage by short-reads in METs. **D3)** Distance between the Transcription Termination Sites (TTS) of the METs and their FSM references. Same MET means both GlR and ExR select the same MET, Different MET means GlR selects a MET that is not manually curated and ExR selects a MET that is manually curated.

Next, we evaluated the accuracy in transcript quantification of the two above mapping strategies, and compared them with quantification solely based on PacBio data considering the count of FL reads of the sequenced transcripts.

Accuracy was evaluated as correlation between replicates of the same cell type. To account for the influence that different noise levels at low and highly expressed transcripts could have on correlation, we computed values binning for high, medium and low expressed transcripts. Although generally high when based on short-read data, correlations in the medium and low expression ranges were higher for ExR than GlR based quantifications (Figure 6B), suggesting that using a reduced transcriptome as reference for short-read mapping improves expression estimates. In contrast, quantification based on PacBio Full-length reads was significantly lower for the same samples, especially at the medium and lower expression ranges where the values dropped to nearly zero (Figure 6B). This shows that, at least at the current sequencing depth of PacBio, the number of reads would not be sufficient for an accurate quantification of transcript expression.

One motivation for a transcript-resolved analysis is the identification of the transcript that captures most of the expression in each gene. We analyzed the concordance between the ExR and GlR quantification selecting the Most Expressed Transcript (MET) of the gene. For 3,930 genes the MET was identical in both the ExR and GlR quantifications, meaning that the ExR MET was a Full Splice Match (FSM) of the GlR MET, while the ExR and GlR METs were different for 1,092 genes. For example, the Dehydrogenase/reductase member 7B (*Dhrs7b*) gene is expressed as two transcripts in our PacBio neural transcriptome (PB.1035.1, PB.1035.2) and had three transcripts in RefSeq quantification (NM_001172112, NM_145428 and XM_006532689) (Figure 6C.1). PB.1035.1 is a FSM transcript of NM_145428 and was the most expressed transcript according to the ExR quantification (Figure 6C.2), while PB.1035.2 is a FSM transcript of NM_001172112. Although both PacBio transcripts have exact junction matches to their respective RefSeq FSMs, they have shorter 3' exons (Figure 6C.1). This shorter 3' exon is actually the annotated exon of a third RefSeq transcript, XM_006532689, which also uses an alternative 5' exon (Figure 6C.1). XM_006532689 was the MET in the GlR quantification while the other two transcripts were estimated as poorly expressed (Figure 6C.2). Upon RT-PCR amplification with transcript discriminating primers we confirmed the ExR and not the GlR based quantification scheme (Figure 6C.3). When inspecting read coverage at this locus we observed that neither the unique 5' junction of XM_006532689 nor the extra exonic sequence at the 3'exon of NM_001172112 or NM_145428 were covered by Illumina short reads, while the short-read pattern nicely fits the PacBio transcript models (Figure 6C.1, detailed in red dashed boxes). We speculate that this variability at the 3'ends creates a conflict when resolving transcript quantification in the RefSeq gene model that was decided in favor of transcript XM_006532689 by our quantification algorithm (RSEM[41]), as this transcript has a more consistent 3' end coverage. In summary, the transcript quantification error of the *Dhrs7b* gene when using a reference transcriptome as mapping template was due to a discrepancy in the 3'end annotation between the reference and the actual expressed transcripts. Similar disagreement patterns were observed for two additional genes, *Spcs2* and *Bdkrb2* with similar outcomes in terms of MET selection (Supplementary Figure 7).

To estimate how general this pattern was, for all the MET discrepant genes we investigated the RefSeq curation status. Interestingly, the majority of the discrepant genes (54%, n=588 genes) corresponded to situations where the ExR-based MET was a FSM of a manually curated RefSeq transcript and the GlR-based MET was not manually curated, as in the *Dhrs7b* gene. Furthermore, in these cases the GlR-based MET had significantly worse lowest

splice junction coverage and lowest mean exon coverage than the MET called by the ExR quantification (Figures 6D.1 and 6D.2). Similarly to *Dhrs7b*, we found that for these 588 genes the differences in the length at the 3' end between the MET selected at ExR quantification and its matched RefSeq transcript were significantly higher than in genes where both quantifications selected equivalent METs (Figure 6D.3). Moreover, these differences were also observed for transcripts codifying for the PI-ORF of the genes, indicating that the extensive variability in the 3' ends that is not annotated in a global reference such as RefSeq is not restricted to secondary/alternative transcripts. Taken together these results demonstrate that providing an expressed, full-length transcriptome as reference for short read quantification improves gene expression estimations, not only because of their reduced complexity, but also due to differences between the reference gene model annotations and the actually expressed transcripts, especially at the more variable 3' ends.

DISCUSSION

SQANTI as a critical tool to analyze whole transcriptome quality.

Long read sequencing technologies, such as the PacBio platforms as well as Illumina's Moleculo and Oxford Nanopore, have brought novel excitement into the challenge of describing the complexity of the transcriptome of higher eukaryotes by providing new means for sequencing full-length transcript models. While early papers concentrated on demonstrating the dramatic enrichment in full-length transcripts achieved by long reads[23,42], there is an increasing number of publications that describe thousands of new transcripts discovered by this technology. Accordingly, we found that, when sequencing the mouse neural transcriptome using PacBio a large number of novel transcripts could be detected. However, close inspection of these new transcripts revealed signs of potential errors that required a thorough and systematic analysis of these sequences before making any new transcript calls. This motivated the development of SQANTI, a new software for the structural and quality analysis of transcripts obtained by long-read sequencing.

The three basic aspects of the SQANTI pipeline are i) the classification of transcripts according to the comparison of their junctions to a reference annotation in order to dissect the origin of transcript diversity, ii) the computation of 32 different descriptors to chart transcript characteristics and iii) the generation of graphs from descriptors data, frequently with a transcript-type break-down, to facilitate interpretation of the sequencing output and reveal potential biases in the novel sequences. Using this analysis framework we were able to show that, at least in our mouse experiment, novel transcripts - especially those in the NNC category - are typically poorly expressed transcripts of known genes and that novel junctions accumulate at the 5' end of transcripts, have lower coverage by Illumina reads, and are enriched in non-canonical splicing and direct repeats typical of RT-switching. However, none of these features are exclusive of any of the novel transcripts categories, which invites the question on "how best to remove transcript artifacts". This has been solved in the past by either eliminating all novel transcripts with at least one junction not supported by short-reads[23], by systematically discarding transcripts with non-canonical splicing[26] or by developing models to estimate the likelihood of a certain splicing event[28]. In our case, we performed an extensive PCR validation of transcripts belonging to different known and novel types. Surprisingly, we found a significant number of transcripts, both with canonical and non-canonical junctions, that did have complete junction

support by Illumina and were amplified by RT-PCR of the sequenced cDNA library, but failed to be validated when PCR conditions were adjusted to avoid secondary RNA structures. We concluded that these might be cases of retrotranscription artifacts, which would have escaped a filtering solely based on short-read support. This result may suggest that a revision of library preparation protocols is needed, which goes beyond the scope of this study. As an alternative, we were able to combine our set of SQANTI descriptors with a machine learning strategy to build a filter that discards poor quality transcripts with higher accuracy than the methods indicated above. We do not claim our current filter is universally valid, but we believe that the selected features and the SQANTI analysis framework can be extensively used by everyone to assess data quality and make educated decisions when preprocessing PacBio data.

Our long-read sequencing quality survey goes beyond the SQANTI analysis to also interrogate aspects related to quantification of transcript expression and translation of alternative transcripts. Previous work showed that using a reduced transcriptome as reference for short-read mapping improves accuracy in the estimation of gene expression [26,39]. We confirmed these findings, but more interestingly, we show how high variability at transcript ends is a source of quantification errors that can be alleviated when an expressed full-length reference transcriptome is used. Our data suggests that unannotated alternative polyadenylation events are frequent in mammalian genomes, which in turn induce incorrect quantification estimates. Full-length sequencing of the expressed transcriptome readily identifies this 3'end diversity to provide the correct templates for transcript quantification. On the other hand, variability at the 5' end is still an issue for full-length transcriptome sequencing as biological variability cannot be unequivocally differentiated from technical artifacts in cDNA library preparation protocols. The SMARTer protocol typically used in PacBio sequencing may not always capture the full extension of the 5' ends due to transcript degradation or incomplete retrotranscription. This may account for the lack of 5' end coverage observed in FSM, ISM and UTR3 Fragment transcripts. Interestingly, trapping of the 5' CAP prior to the synthesis of the secondary cDNA strand has recently been shown to increase the overlap of the 5' end without seriously compromising the yield of long reads[43] and in future may represent the preferred form of library preparation to study 5' end diversity.

Finally, we investigated whether the transcriptome diversity found by long read sequencing was mirrored by proteogenomics data. We concluded that the low expression and identifiability by single peptides of Alt and Novel ORFs hampered their detection by proteomics. Detection of alternative protein isoforms has proved to be difficult, and while some authors claim that limited detection in proteomics databases indicates low translational and stability rates[44,45], other studies identify a significant proportion of alternative exons associated to ribosomes as evidence of active translation[46,47]. While it is not the scope of this work to resolve these issues, we turned our attention to the analysis of protein differences for those cases of confident peptide detection. Interestingly, we found that the distribution of the type of protein differences in the non PI-ORFs with respect to the main isoforms is similar to the predictions based on the PacBio sequencing data, except for N-terminal truncations that are at a disadvantage in a unique peptide detection approach. Most of detected alternative ORFs showed major protein changes compared to the PI-ORF of their respective genes, which could potentially have an impact on functionality of the alternative protein. While a detailed analysis of these functional differences requires further computational and experimental

approaches, the results presented in this paper indicate that long read technologies, provided adequate quality control is applied, are effective tools for describing the isoform-resolved transcriptome and can aid in the study of the biological significance of alternative splicing.

MATERIAL AND METHODS

Differentiation of NPCs and OLPs from neonatal mice

Neonatal c57/BL6 mice (4 days old) were sacrificed and Neural Precursors cells (NPCs) were isolated from the subventricular zone. Neurospheres were obtained by culturing the progenitors in media supplemented with EGF and bFGF and Oligodendrocyte Precursor Cells (OPCs) were derived from them by adding ATRA (All Trans Retinoic Acid) as described in the Supplementary Methods section.

RNA extraction, full-length cDNA library preparation and sequencing.

Total RNA isolation from cultured cells (two biological replicas per cell type) was done with Nucleospin RNA kit (Macherey-Nagel) obtaining RINs (RNA Integrity Number) between 10 and 9.7 for all samples. The synthesis of full-length cDNA was performed with SMARTer PCR cDNA Synthesis kit (ClonTech, CA, USA) following PacBio recommendations. The cDNA synthesis protocol used 1 µg of total RNA, 42 ºC for retrotranscription and 13 PCR amplification cycles to control for over-amplification of small fragments. For each sample, we performed two first-strand cDNA synthesis reactions and nine PCR reactions using 10 µl of first strand cDNA (diluted 1:5 in TE-buffer) to obtain around 14-16 µg full-length cDNA per sample. Each sample was submitted to the ICBR sequencing facility (University of Florida) for PacBio sequencing. Three cDNA fractions were obtained with BluePippin and sequenced at the RSII Instruments using 2 SMRT cells for the 1-2 kb fraction, and 3 SMRT cells for 2-3 kb and 3-6 kb fractions, to a total of 8 SMRT cells per sample. Sequenced PacBio subreads were pooled together and ToFU software was used to obtain non-redundant transcripts. Default parameters were set to obtain Read of Insert (RoI), Full-length classification of RoIs and ICE (Iterative Clustering for Error Correction) steps. Quiver option was turned on to improve consensus accuracy of previously generated ICE clusters by using non Full Length read information. Finally, generated HQ polished isoforms (>99 % accuracy after polishing) were collapsed to eliminate isoform redundancy (5' different was not considered when collapsing isoforms). This set of 5' merged non-redundant isoforms was defined as ToFU transcriptome. Additionally, the same samples were sequenced with the Illumina Nextseq instrument using Nextera tagmentation and 2x50 paired end sequencing. Further details can be found in the Supplementary Methods section.

The PacBio raw reads have been submitted to the SRA under Submission number SUB2459157 (PacBio reads) and SUB2466432 (Illumina reads) will be released upon publication of this manuscript.

Verification of transcripts by Reverse Transcription PCR (RT-PCR)

PCR amplification of selected transcripts was performed with both the sequenced full-length cDNA and newly synthesized cDNA from the same RNA extractions. For new cDNA reactions, 1 µg of total RNA was used to synthesize the first-strand cDNA using SuperScript III (Life Technologies) primed with random hexamers in a

reaction volume of 20 µl, according to the manufacturer's instructions. Each random hexamer cDNA synthesis reaction was carried out at two temperature conditions: 42 ºC and 50 ºC. RT-PCR reactions used 1 µl of sequenced full-length cDNA or 2 µl of random hexamers cDNA, together with Biotools DNA Polymerase (1U/ µl) in a reaction volume of 50 µl. Primers were designed to span the predicted splicing event using Primer-BLAST[48] Supplementary Table 3, http://www.ncbi.nlm.nih.gov/tools/primer-blast). PCR condition were 5 min at 94 ºC followed by 35 cycles of 94 ºC 30 s, primer-specific annealing temperature for 30 s and 72 ºC for 1 min or 1:30 min, depending of predicted product size. PCR amplification was monitored on 1.5 % agarose gel.

SQANTI pipeline

The SQANTI pipeline is implemented in Python with calls to R for generation of transcriptome descriptive plots. The SQANTI function performs different tasks: (1) corrects transcript sequences based on the provided reference and returns a corrected transcriptome; (2) compares sequenced isoforms with current genome annotation to generate genes models and classify transcripts according to splice junctions (a full-description of structural classification of isoforms can be found in results section); (3) predicts ORFs using GeneMarkST (4) runs our algorithm to predict RT switching (5) annotates transcripts and junction according to SQANTI descriptors listed in Supplementary Table 1. SQANTI has been released accompanied by a function to filter out transcripts that are likely to be artifacts, which takes as input the output generated by the main SQANTI function and two sets of true positive and true negative transcripts defined by the user. SQANTI filtering generates a classifier of transcripts and provides a curated transcriptome where artifactual transcripts are removed. SQANTI is available at https://bitbucket.org/ConesaLab/sqanti.

RT switching prediction

SQANTI contains an algorithm that implements the RT switching (RTS) conditions described in Cocquet et al[36]. Namely an exon skipping pattern due to a retrotranscription gap caused by secondary structures in expressed transcripts. The algorithm looks at all the junctions for possible RTS (both canonical and non-canonical junctions) and checks for a direct repeat pattern match at defined sequence locations: the pattern at the end of the splice junction's 5' exon must match the pattern at the 3' end of the splice junction's intron. There are three parameters that control pattern matching: (1) the minimum number of nts required to match (4 - 10); (2) the number of nts of wiggle allowed from the ideal pattern location (0 - 3); (3) whether allow for a single mismatch, indels or not. SQANTI uses as default parameters: a minimum of 8 bases long repeat sequences, a maximum wiggle of 1 and no mismatches. FSM transcripts with the highest mean expression in each gene are assumed to serve as templates for RTS and are excluded from the analysis.

ORF prediction with GMST within SQANTI

The GMST algorithm[13] was applied to predict ORFs in PacBio transcripts, setting parameters to only consider the direct strand of the cDNA and AUGs as the initial codon. As GeneMarkS-T allows prediction in incomplete transcripts lack of coverage in the 5' end caused some truncated ORF starting in codons different from Methionine.

In these instances the ORF was shortened by the N-Terminus until the first in frame Methionine was found. GMST was benchmarked as shown in Supplementary methods.

Quality Control filtering based on SQANTI features

A machine learning approach was developed to discriminate artifacts from true novel transcripts based on SQANTI features. We defined as training set the transcripts analyzed by PCR: 22 PCR positives were considered true transcripts and 35 PCR negatives were considered artifacts. We pre-selected a set of 19 SQANTI descriptors (Supplementary Table 1) as input variables of the predictor. These variables were reduced to 16 after discarding highly correlated descriptors. Several machine learning methods were tested (Adaboost[49], CART[50], Random Forest[37], SVM[51], Treebag[52]) and Random Forest was selected as best performing approach (Supplementary Methods). The Random Forest algorithm was applied using 250 trees, 2 variables assessed at each split, down sampling and 10 times 10 cross validation to evaluate performance. The resulting classifier was applied to all novel transcripts to filter out false novel calls.

Analysis of Peptide Support

We performed an *in silico* analysis of the peptide support of the predicted ORFs of our neural transcriptome when compared to public proteomics databases. A non-redundant database composed of predicted ORFs from our murine transcriptome experiments and all the murine ORFs annotated in Ensembl (v80) was created. These ORFs were subjected *in silico* tryptic digestion (Proteogest, complete digestion). Unique peptides were identified and ORFs with at least one unique peptide of 7 amino acids of more were annotated as *identifiable ORFs*. We then used two different approaches to detect experimental Peptide to Spectrum Matches (PSMs) that match unique peptides from our ORFs. The first approach made use of a pipeline built on Pladipus[53], a platform that allows for distributed and automated execution of bio-informatics related tasks and performed an all tissue search of mouse proteomic studies (n=36). The pipeline consists of pride-asap, a tool designed to automatically extract optimal search parameters, SearchGUI[54], a tool that manages the execution of several search engines and PeptideShaker[55], a tool that allows for the merging of the results produced by the search engines. For this study, X!Tandem[56], Myrimatch[57] and MSGF+[58] algorithms were applied. The input spectra were obtained from 36 murine projects in the PRIDE[59] database. The second approach was based on the Sequest algorithm[60] and screened large-scale mouse proteomics experiments of brain tissue[61] and astrocyte secreted proteins[62]. A more detailed description of these approaches is available in Supplemental Methods.

Characterization of Alt-ORF and Novel-ORF with respect to PI-ORFs

Microexon definition was restricted to novel amino-acid (aa) stretches obtained by in-frame indels or substitutions of no more than 27 nts (9aas) following Irimia et al[63]. ORFs showing exclusively N-Ter deletions or C-Ter deletions were labeled as N-Ter Deletion or C-Ter Deletion ORFs. ORFs showing indels and substitutions greater than 9 aas, combined or not with N-Ter and C-Ter deletions, were labeled as Major Change ORFs. ORFs that could not be aligned against the PI-ORF of their respective genes were deemed as No align ORFs.

Transcript quantification

Transcript quantification using short-reads were obtained with STAR[64] as mapper and RSEM[41] as quantification algorithm (parameters available at Supplementary Methods), using as reference transcriptome the gtf files of the corrected *PacBio transcriptome* (16,104 transcripts), the *curated transcriptome* (after SQANTI filtering, 12,404) and the RefSeq database (~116,000 transcripts). Expression estimates were obtained as Transcript per million (TPM). Long-read quantification was computed as the number of Full-length reads of each transcript divided by the total number of FLs of the sample. To compute correlation among duplicates of the same condition, expression values of both replicates were first averaged and the 1st and 3rd quartile of the mean expression value distribution was used to bin transcripts as low (from minimum to 1st quartile), medium (from 1st to 3rd quartile) and highly (from 3rd to maximum) expressed. Then, correlation between replicates was calculated within each bin. Most Expressed Transcript (MET) was defined as the transcript of each gene that obtained the highest average TPM value across all the samples.

ACKNOWLEDGEMENTS

We thank Eric Triplett (University of Florida) for support in sequencing experiments and Elizabeth Tseng (PacBio) for helping in running the ToFU pipeline and critically reading this manuscript. This work has been partially funded by the University of Florida Pre-eminence hires program, the Spanish Ministry of Economy and Competitiveness grant BIO2015-71658-R and Spanish Ministry of Education grant FPU2013/02348.

REFERENCES

1. Frankish, A., Mudge, J. M., Thomas, M. & Harrow, J. The importance of identifying alternative splicing in vertebrate genome annotation. *Database* **2012,** (2012).

2. Lu, T. *et al.* Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. *Genome Res.* **20,** 1238–1249 (2010).

3. Mudge, J. M. *et al.* The origins, evolution, and functional potential of alternative splicing in vertebrates. *Mol. Biol. Evol.* **28,** 2949–2959 (2011).

4. Martinez, N. M. & Lynch, K. W. Control of alternative splicing in immune responses: Many regulators, many predictions, much still to learn. *Immunological Reviews* **253,** 216–236 (2013).

5. Raj, B. & Blencowe, B. J. Alternative Splicing in the Mammalian Nervous System: Recent Insights into Mechanisms and Functional Roles. *Neuron* **87,** 14–27 (2015).

6. Teichroeb, J. H., Kim, J. & Betts, D. H. The Role of Telomeres and Telomerase Reverse Transcriptase Isoforms in Pluripotency Induction and Maintenance. *RNA Biol.* **6286,** 00–00 (2016).

7. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10,** 57–63 (2009).

8. McGuire, A. M., Pearson, M. D., Neafsey, D. E. & Galagan, J. E. Cross-kingdom patterns of alternative splicing and splice recognition. *Genome Biol.* **9,** R50 (2008).

9. Chen, J. & Weiss, W. a. Alternative splicing in cancer: implications for biology and therapy. *Oncogene* **34,** 1–14 (2014).

10. Ladomery, M. Aberrant alternative splicing is another hallmark of cancer. *International Journal of Cell Biology* (2013). doi:10.1155/2013/463786

11. Liu, S. & Cheng, C. Alternative RNA splicing and cancer. *Wiley Interdisciplinary Reviews: RNA* **4,** 547–566 (2013).

12. Eizirik, D. L. *et al.* The human pancreatic islet transcriptome: Expression of candidate genes for type 1 diabetes and the impact of pro-inflammatory cytokines. *PLoS Genet.* **8,** (2012).

13. Tang, S., Lomsadze, A. & Borodovsky, M. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res.* **248,** 1–18 (2015).

14. D'Souza, I. *et al.* Missense and silent tau gene mutations cause frontotemporal dementia with parkinsonism-chromosome 17 type, by affecting multiple alternative RNA splicing regulatory elements. *Proc. Natl. Acad. Sci. U. S. A.* **96,** 5598–603 (1999).

15. Kanadia, R. N. *et al.* A muscleblind knockout model for myotonic dystrophy. *Science* **302,** 1978–80 (2003).

16. Ladd, A. N. CUG-BP, Elav-like family (CELF)-mediated alternative splicing regulation in the brain during health and disease. *Mol. Cell. Neurosci.* **56,** 456–464 (2013).

17. Lee, J. A. *et al.* Cytoplasmic Rbfox1 Regulates the Expression of Synaptic and Autism-Related Genes. *Neuron* **89,** 113–128 (2016).

18. Yang, Y. Y., Yin, G. L. & Darnell, R. B. The neuronal RNA-binding protein Nova-2 is implicated as the autoantigen targeted in POMA patients with dementia. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 13254–9 (1998).

19. Black, D. L. Mechanisms of Alternative Pre-Messenger RNA Splicing. *Annu. Rev. Biochem.* **72,** 291–336 (2003).

20. Cognata, V. La *et al.* Increasing the Coding Potential of Genomes Through Alternative Splicing: The Case of PARK2 Gene. *Curr. Genomics* **15,** 203–16 (2014).

21. Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* **10,** 1177–84 (2013).

22. Tilgner, H., Grubert, F., Sharon, D. & Snyder, M. P. Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proc. Natl. Acad. Sci. U. S. A.* **1640,** 10–12 (2014).

23. Sharon, D., Tilgner, H., Grubert, F. & Snyder, M. A single-molecule long-read survey of the human transcriptome. *Nat. Biotechnol.* **31,** 1009–14 (2013).

24. Song Li, Masashi Yamada, Xinwei Han, Uwe Ohler, P. N. B. High-Resolution Expression Map of the Arabidopsis Root Reveals Alternative Splicing and lincRNA Regulation. *Dev. Cell* (2016).

25. Tilgner, H. *et al.* Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nat. Biotechnol.* **33,** 736–742 (2015).

26. Au, K. F. *et al.* Characterization of the human ESC transcriptome by hybrid sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **110,** E4821--30 (2013).

27. Au, K. F., Underwood, J. G., Lee, L. & Wong, W. H. LSC Improving PacBio Long Read Accuracy by Short Read Alignment. *PLoS One* **7,** 1–8 (2012).

28. Abdel-Ghany, S. E. *et al.* A survey of the sorghum transcriptome using single-molecule long reads. TL - 7. *Nat. Commun.* **7 VN-re,** 11706 (2016).

29. Gordon, S. P. *et al.* Widespread polycistronic transcripts in fungi revealed by single-molecule mRNA sequencing. *PLoS One* **10,** (2015).

30. Wang, B. *et al.* Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nat. Commun.* **7,** 11708 (2016).

31. Rogers, M. F., Thomas, J., Reddy, A. S. & Ben-Hur, A. SpliceGrapher: detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data. *Genome Biol.* **13,** R4 (2012).

32. Wu, T. D. & Watanabe, C. K. GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21,** 1859–1875 (2005).

33. Hackl, T., Hedrich, R., Schultz, J. & Förster, F. proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics* **30,** 1–8 (2014).

34. Rodriguez, J. M. *et al.* APPRIS: Annotation of principal and alternative splice isoforms. *Nucleic Acids Res.* **41,** (2013).

35. Parada, G. E., Munita, R., Cerda, C. A. & Gysling, K. A comprehensive survey of non-canonical splice sites in the human transcriptome. **42,** 10564–10578 (2014).

36. Cocquet, J., Chong, A., Zhang, G. & Veitia, R. A. Reverse transcriptase template switching and false alternative transcripts. *Genomics* **88,** 127–131 (2006).

37. Breiman, L. Random Forests. *Mach. Learn.* **45,** 5–32 (2001).

38. Deutsch, E. W. *et al.* Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 2.1.

Journal of Proteome Research **15,** 3961–3970 (2016).

39. Soneson, C., Matthes, K. L., Nowicka, M., Law, C. W. & Robinson, M. D. Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage. *Genome Biol.* **17,** 12 (2016).

40. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).

41. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12,** 323 (2011).

42. Tseng, E. & Underwood, J. G. Full Length cDNA Sequencing on the PacBio RS. in *ABRF* **24,** (2013).

43. Cartolano, M., Huettel, B., Hartwig, B., Reinhardt, R. & Schneeberger, K. cDNA library enrichment of full length transcripts for SMRT long read sequencing. *PLoS One* **11,** (2016).

44. Tress, M. L., Abascal, F. & Valencia, A. Alternative Splicing May Not Be the Key to Proteome Complexity. *Trends in Biochemical Sciences* (2016). doi:10.1016/j.tibs.2016.08.008

45. Ezkurdia, I. *et al.* Most highly expressed protein-coding genes have a single dominant isoform. *J. Proteome Res.* **14,** 1880–1887 (2015).

46. Weatheritt, R. J., Sterne-Weiler, T. & Blencowe, B. J. The ribosome-engaged landscape of alternative splicing. *Nat. Struct. Mol. Biol.* 1–9 (2016). doi:10.1038/nsmb.3317

47. Sterne-Weiler, T. *et al.* Frac-seq reveals isoform-specific recruitment to polyribosomes. *Genome Res.* **23,** 1615–1623 (2013).

48. Ye, J. *et al.* Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* **13,** 134 (2012).

49. Schwenk, H. & Bengio, Y. Boosting neural networks. *Neural Comput.* **12,** 1869–1887 (2000).

50. Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. *Classification and Regression Trees*. *The Wadsworth statisticsprobability series* **19,** (1984).

51. Cortes, C. & Vapnik, V. Support-Vector Networks. *Mach. Learn.* **20,** 273–297 (1995).

52. Loh, W.-Y. & Shih, Y.-S. Split Selection Methods for Classification Trees. *Stat. Sin.* **7,** 815–840 (1997).

53. Verheggen, K. *et al.* Pladipus Enables Universal Distributed Computing in Proteomics Bioinformatics. *J. Proteome Res.* **15,** 707–712 (2016).

54. Vaudel, M., Barsnes, H., Berven, F. S., Sickmann, A. & Martens, L. SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *Proteomics* **11,** 996–999 (2011).

55. Vaudel, M. *et al.* PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat. Biotechnol.* **33,** 22–24 (2015).

56. Craig, R. & Beavis, R. C. TANDEM: Matching proteins with tandem mass spectra. *Bioinformatics* **20,** 1466–1467 (2004).

57. Tabb, D. L., Fernando, C. G. & Chambers, M. C. MyriMatch: Highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J. Proteome Res.* **6,** 654–661 (2007).

58. Kim, S., Gupta, N. & Pevzner, P. A. Spectral probabilities and generating functions of tandem mass spectra: A strike against decoy databases. *J. Proteome Res.* **7,** 3354–3363 (2008).

59. Martens, L. *et al.* PRIDE: The proteomics identifications database. *Proteomics* **5,** 3537–3545 (2005).

60. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* **5,** 976–989 (1994).

61. Sharma, K. *et al.* Cell type- and brain region-resolved mouse brain proteome. *Nat. Neurosci.* **18,** 1819–1831 (2015).

62. Han, D., Jin, J., Woo, J., Min, H. & Kim, Y. Proteomic analysis of mouse astrocytes and their secretome by a combination of FASP and StageTip-based, high pH, reversed-phase fractionation. *Proteomics* **14,** 1604–1609 (2014).

63. Irimia, M. *et al.* A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* **159,** 1511–1523 (2014).

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