nature biotechnology

Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells

Ishaan Gupta[1,9], Paul G Collier[1,9], Bettina Haase[2], Ahmed Mahfouz[1,3,4], Anoushka Joglekar[1], Taylor Floyd[1], Frank Koopmans[5], Ben Barres[6,8], August B Smit[5], Steven A Sloan[6], Wenjie Luo[7], Olivier Fedrigo[2], M Elizabeth Ross[1] & Hagen U Tilgner[1]

Full-length RNA sequencing (RNA-Seq) has been applied to bulk tissue, cell lines and sorted cells to characterize transcriptomes[1–11], but applying this technology to single cells has proven to be difficult, with less than ten single-cell transcriptomes having been analyzed thus far[12,13]. Although single splicing events have been described for ≤200 single cells with statistical confidence[14,15], full-length mRNA analyses for hundreds of cells have not been reported. Single-cell short-read 3′ sequencing enables the identification of cellular subtypes[16–21], but full-length mRNA isoforms for these cell types cannot be profiled. We developed a method that starts with bulk tissue and identifies single-cell types and their full-length RNA isoforms without fluorescence-activated cell sorting. Using single-cell isoform RNA-Seq (ScISOr-Seq), we identified RNA isoforms in neurons, astrocytes, microglia, and cell subtypes such as Purkinje and Granule cells, and cell-type-specific combination patterns of distant splice sites[6–9,22,23]. We used ScISOr-Seq to improve genome annotation in mouse Gencode version 10 by determining the cell-type-specific expression of 18,173 known and 16,872 novel isoforms.

Unlike sorting-based methods (**Supplementary Fig. 1a**), ScISOr-Seq identifies isoforms in >1,000 single cells from bulk tissue without cell sorting by combining two technologies (**Fig. 1a**). We used microfluidics to amplify full-length cDNA from single cells in a sample. cDNA produced from each single cell was barcoded to enable cell-of-origin identification and then split into two pools, with one pool being used for short-read Illumina 3′ sequencing to measure gene expression and the other pool being used for long-read sequencing and isoform identification. Short-read 3′ sequencing provided molecular counts for each gene and cell, which enabled clustering of cells and cell type assignment using cell-type-specific markers. Long-read sequencing with Pacific Biosciences (PacBio)[1,2,4,5] or Oxford Nanopore[3] was used to identify full-length RNA isoforms. Single-cell barcodes were also present in long reads and could be used to determine the individual cell of origin for each long read. Given that most single cells are assigned to a named cluster, we were also able to assign the cluster name, for example, 'Purkinje cell' or 'astrocyte', to each long read (**Fig. 1a** and Online Methods).

We used ScISOr-Seq to identify cell-type-specific isoforms in mouse cerebellum at postnatal day 1 (P1). We sequenced a mean of 17,885 reads per cell (according to 10xGenomics' summary statistics). After filtering cells (Online Methods) to retain reads confidently mapped to genes, we had 3,875 unique molecular identifiers (UMIs) and 1,448 genes per cell (**Supplementary Fig. 2a–d**). We used these short reads to cluster 6,627 cells into 17 groups (**Fig. 1b**, **Supplementary Fig. 2d**, Online Methods and **Supplementary Code**). High expression of well-established cell-type-specific markers identified many clusters as cell types. High expression of *Pdgfra*, *Olig1* and *Olig2* identified a cluster of oligodendrocyte precursors (OPCs; **Fig. 1b,c**). *Clu* and *Apoe* identified two clusters of astrocytes and *Gdf10* (refs. 24,25) identified a cluster of Bergmann glia (BG). We also identified three large clusters of neuronal subtypes: the external granular layer (EGL) cell cluster, marked by expression of *Neurod1* and *Ccdn2*, contained cells in several stages of differentiation; Purkinje cells, marked by expression of *Pcp4*, *Gad1* and *Gad2* in the Purkinje cell layer (PCL); and other neurons known to be present in the deep cerebellar nuclei (DCN) cluster close to internal granular layer (IGL) neurons. Together, DCN and IGL neurons expressed *Pnoc*, *Snhg11*, *Tcf7l2*, *Gad1*, *Gad2* and *Lhx9*. The proximity of DCN and IGL neurons in clustering probably reflects their overlapping embryonic origins. Given this proximity of DCN and IGL, and the smaller number of long reads for both clusters when separated, we grouped these clusters and collectively refer to these two populations as IGL-DCN (**Fig. 1b**). This should not be interpreted as DCN and IGL being identical. These cell-type-specific expression patterns exhibited specific anatomical localization in the developing cerebellum (**Fig. 1c**)[26]. Three additional clusters, each representing between 2–5% of all cells, expressed genes associated with neural progenitor cells, including *Ccnd2* (which is highly expressed in the postnatal EGL), *Atoh1* (glutamatergic neuron precursors from the rhombic

[1]Brain and Mind Research Institute and Center for Neurogenetics, Weill Cornell Medicine, New York, New York, USA. [2]The Rockefeller University, New York, New York, USA. [3]Leiden Computational Biology Center, Leiden University Medical Center, Leiden, the Netherlands. [4]Delft Bioinformatics Lab, Delft University of Technology, Delft, the Netherlands. [5]Department of Molecular and Cellular Neurobiology, Center for Neurogenomics and Cognitive Research, Amsterdam Neuroscience, VU University, Amsterdam, the Netherlands. [6]Department of Neurobiology, Stanford University, Stanford, California, USA. [7]Brain and Mind Research Institute and Appel Alzheimer's Research Institute, Weill Cornell Medicine, New York, New York, USA. [8]Deceased. [9]These authors contributed equally to this work. Correspondence should be addressed to H.U.T. (hut2006@med.cornell.edu).

Received 2 January; accepted 20 August; published online 15 October 2018; doi:10.1038/nbt.4259

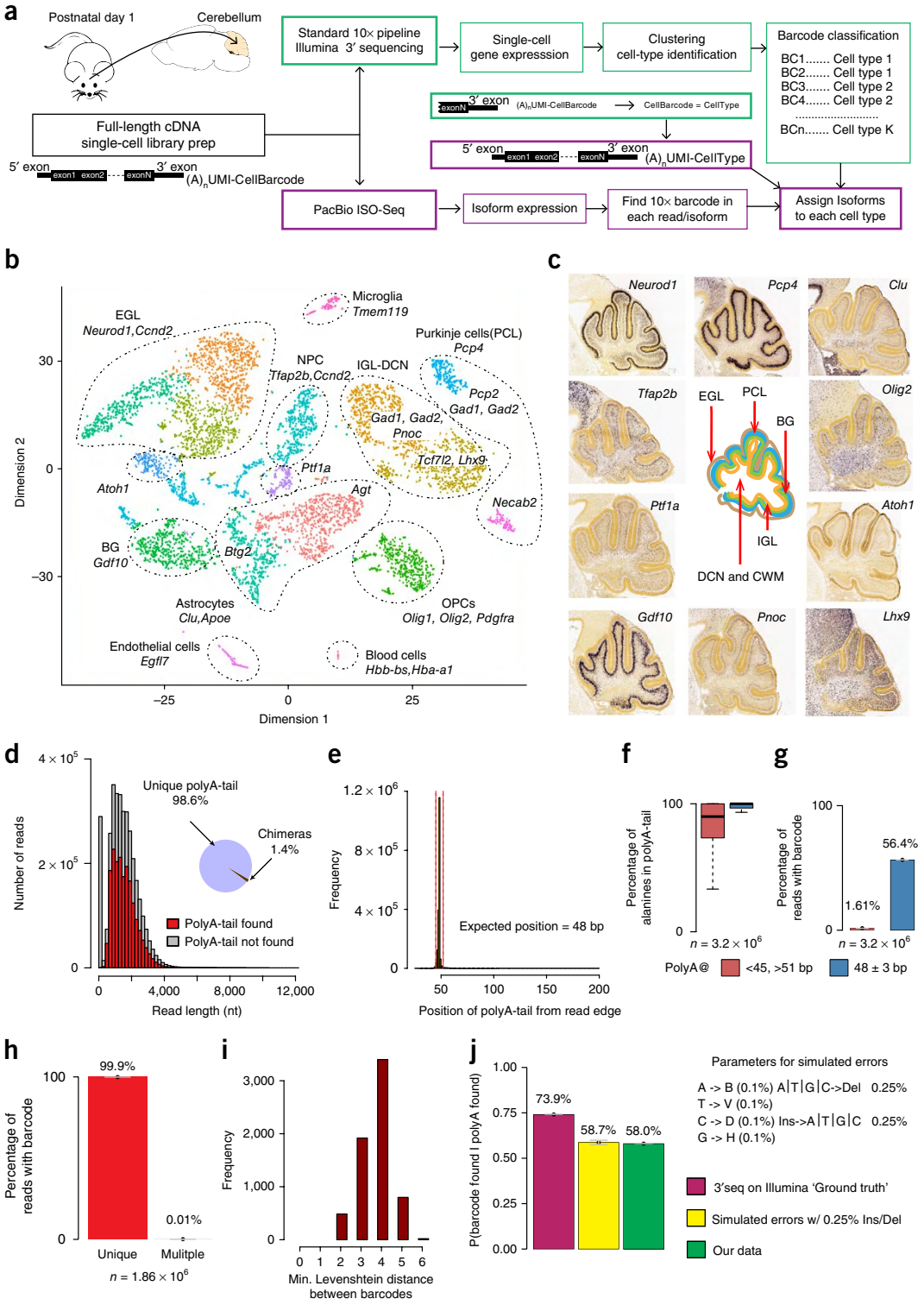

Figure 1 Outline of approach, cell-type and barcode identification. (**a**) Outline of ScISOr-Seq strategy. (**b**) T-distributed stochastic neighbor embedding plot depicting cell clusters, marker genes and names given to clusters, including BG, EGL, IGL-DCN, two clusters of PCL, OPCs, Atoh1^{+} neuronal progenitors, Ptf1a^{+} neuronal progenitors and other neuronal progenitors (NPCs). (**c**) *In situ* hybridization images from the Allen Developing Mouse Brain Atlas showing expression of marker genes in specific layers (image credit: Allen Institute). (**d**) Length distribution of CCS with and without polyA-tail and pie chart, giving the relative abundance of CCS that had exactly one or multiple polyA-tails. (**e**) Histogram of start position of first occurrence of nine consecutive threonines. (**f**) Percent of alanines in polyA-tails for CCS having a T9 between positions 45 and 51 and CCS having a polyA-tail outside these regions. (**g**) Percentage of reads having a barcode, whose T9 starts between positions 45 and 51 and outside of that. Whiskers represent 95% confidence intervals. (**h**) Percentage of reads with a perfect-match barcode that had exactly one such barcode and that had multiple barcodes. Whiskers represent 95% confidence intervals. (**i**) For each barcode, we calculated the minimal Levenshtein distance to all other barcodes. Shown is a barplot of these values. Whiskers represent 95% confidence intervals. (**j**) Probability of finding a barcode given the presence of a polyA-tail in our data, using five simulations of errors on 77-mers.

lip and EGL) or *Ptf1a* (GABAergic neuron precursors from the ventricular zone) (**Fig. 1b** and **Supplementary Fig. 1b**). We identified other cell populations: microglia, marked by myeloid-associated genes (for example, *C1qa*, *C1qb*, *C1qc* and *Tmem119)*, and endothelial and circulatory-system cells. Our clustering recapitulates a large proportion of cell types classically observed in P1 cerebellum[27]. EGL, IGL-DCN cells and astrocytic cells were the largest clusters and blood cells were the smallest (**Supplementary Fig. 2e**). Detected reads, short-read UMIs and genes per cell revealed slight differences between cell types, but were of similar orders of magnitude. Consistent with their eventual complexity and maturing extensive arborization, Purkinje cells had the highest number of read, UMI and gene counts, whereas blood cells had the lowest gene count (**Supplementary Fig. 2f–h**).

Sequencing of a second independent replicate (rep2) and within-replicate analyses revealed that all clusters were dissimilar to any other clusters in the same replicate (Jaccard index < 0.34 for all cluster pairs; **Supplementary Fig. 3a,b**). To assess cluster stability, we increased Illumina sequencing depth threefold in rep2. In all of the clusters, 95–100% of cells were attributed to the same cluster with threefold deeper sequencing compared with the original sequencing depth. Comparison of marker genes between clusters in the two replicates using the Jaccard index identified highly similar clusters (**Supplementary Fig. 3c**) with one exception. The smallest cluster (blood cells) in replicate 1 (rep1) was missing from rep2. Cell-type abundance was reproducible between replicates and was highly correlated (Pearson correlation = 0.91, $n = 11$, correlation-test P value $= 4.5 \times 10^{-5}$; **Supplementary Fig. 3d**).

Next, we generated ~5.2 million PacBio circular consensus reads (CCS, PB_rep1; Online Methods). Cellular barcodes are located close to the polyA-tail, so we first searched for polyA-tails. We located the first nine consecutive Ts (T9) in the first 200 bp of each read and its reverse complement. 61.6% of CCS contained a T9, broadly consistent with our previous estimation (67%)[1,4]. Reads with and without T9 had similar lengths, apart from CCS ≤ 200 bp accumulating in non-T9 CCS. 1.4% of T9-CCS had a T9 in the read start and the complement's start. These may include chimeras introduced during reverse transcription, PCR or blunt-end PacBio library preparation (**Fig. 1d**). Error-free sequencing of the theoretical construct (21-bp adaptor sequence, 16-bp cellular barcode, and 10-bp UMI and polyA-tail) yielded a T9 starting at position 48. ~97% of T9-CCS had a T9 starting between positions 45 and 51 (expected-T9-position CCS; **Fig. 1e**). These CCS have almost 100% T-content in a 30-bp window (polydT-primers were 30 nt long) starting at the T9. Non-expected T9-position CCS had lower T-content (**Fig. 1f**). We then searched for perfectly matched 16-mer cellular barcodes between the read start and the polyA-tail (the barcode search region). Expected T9-position CCS showed a higher barcode identification rate than CSS with a T9 in other positions, and 97.2% of CCS with identified barcodes were among the expected T9-position CCS (**Fig. 1g**). For CCS with a perfectly matching 16-mer cellular barcode, 98.8% had exactly one such barcode, and no other barcode had one mismatch with the barcode search region (**Fig. 1h**). In total, for 58.0% (compared with 74.0% for 10x-3′seq) of the polyA-tail-containing CCS, we identified a perfect-match 16-mer cellular barcode to the single cell in which the RNA isoform was transcribed. For all 6,627 barcodes, the minimal editing distance to any other barcode was calculated. For 92.7% of barcodes, this minimal (Levenshtein) distance was 3 or greater, and for the remaining barcodes it was 2. Thus, for most barcodes there was only one specific error pattern (three errors) that would result in a mis-identified cell. However, in most cases, three random errors would discard the read because none of the 6,627 known barcodes would be detected (**Fig. 1i**).

To confirm this hypothesis, we simulated errors (Online Methods) in 42 million 77-mers consisting of 10× read1 adaptor (21 bases), single-cell barcode (16 bases), UMI (10 bases) and a 30-mer polyT-tail (representing the polyA-tail). We detected a false-positive barcode among the 6,627 barcodes in <0.1% of the cases (specificity = 99.99%). However, ~41.3% (average across five simulations) of molecules were discarded, as none of the 6,627 single-cell barcodes were found (sensitivity = 58.7% compared with 74% for 10x-3′seq). To confirm experimentally our high specificity, we synthesized cDNA from the cell line GM12878 with a 25-bp sequencing adaptor, one (fixed for all GM12878 molecules) cellular 16-mer barcode, a 5-bp mock UMI and a polyA-tail. These validation-experiment reads therefore have a ground truth barcode. After 16 cycles of PCR, PacBio sequencing and barcode analysis, we did not find any false-positive barcodes (0 of 88,200), revealing a specificity of $\geq 100 \times [1 - (1/88{,}200)]\% = 99.9989\%$. In summary, experiments and simulations validated the specificity of our single-cell barcode-detection procedure although it does not have perfect sensitivity (**Fig. 1j**). We detected a median of 270 long reads, 260 UMIs and 129 genes per single cell. 3.8% of UMIs were observed twice (the theoretical prediction was 3.4%; Online Methods). 99.3% (6,581 of 6,627) of clustered cells were detected with CCS (**Supplementary Fig. 4a–d**). 97.4% (6,459 of 6,627) of clustered cells had >100 CCS (**Supplementary Fig. 4d**). Detected short-read and long-read UMIs per single cell were highly correlated (Pearson correlation = 0.95, correlation-test $P < 2.2 \times 10^{-16}$; **Supplementary Fig. 4e**). Long-read statistics per cell cluster mirrored those for our short-read data sets (**Supplementary Figs. 2f–h** and **4f–h**), with lower long-read numbers.

We also tested nanopore long-read sequencing using a MinIon R9.5 (Online Methods) and searched for barcodes in 2.3 million Nanopore reads[28]. We found lower relative numbers of '1D' Oxford Nanopore reads with a T9, possibly owing to incorrectly reading homopolymers using a minION[28]. However, ~31.4% (1D) and ~35.2% (passed '1D^2' Oxford Nanopore reads) of nanopore reads had a 30-bp window with ≥25 Ts. Although the variation from the expected position in nanopore reads was larger than for PacBio reads (90 bp versus 3 bp), accumulation around the expected position was observed and exact barcode matches revealed unique barcodes in 6.0% of the passed 1D Oxford Nanopore reads and 32.7% of the passed 1D^2 Oxford Nanopore reads; **Supplementary Fig. 5**). Overall, we were able to detect ~50,000 cluster-specific long reads per flow cell. With each current flow cell requiring 1 µg of cDNA, further PCR (with associated biases) would be needed to carry out large-scale ScISOr-Seq using a minION, whereas only 16 cycles of PCR are needed for 20–50 SMRTcells (PacBio), yielding up to 5 million long reads assigned to single cells. Better performance with a minION would be obtained using longer and more diverse barcodes.

We aligned PacBio reads to the mouse genome[29] (version mm10) using STAR[30] and carried out mapping quality control as described previously[1,4,6] (Online Methods, **Supplementary Note 1** and **Supplementary Fig. 6**). We analyzed novel isoforms with respect to mouse Gencode version 10, as outlined previously[1,6,31] (Online Methods) to produce a long-read-enhanced and cell-type-resolved annotation. We considered 10,691 unique novel (with respect to mouse Gencode version 10) isoforms (Online Methods) that affected 4,859 genes. For these isoforms, we required all splice sites to be known in Gencode[32] (version 10) and each junction and internal exon to be either annotated or observed at least twice in ScISOr-Seq. The unique novel isoforms contained new exon-exon junctions linking previously known splice sites, such as the skipping of exons annotated as constitutive. Artifacts in next-generation sequencing have been demonstrated[33], so to assess whether the long-range 16-cycle PCR in

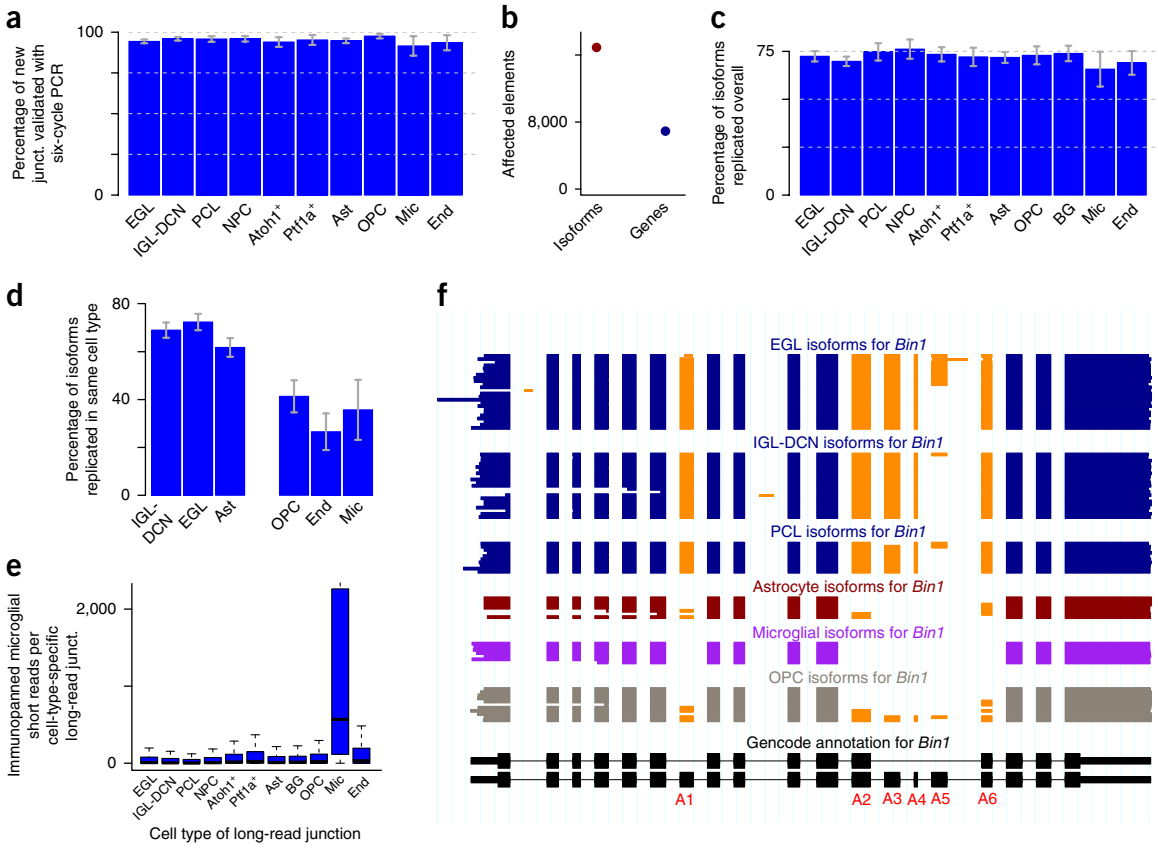

Figure 2 Improved cell-type-specific annotation. (**a**) Percentage of long-read-derived junctions that could be validated using low-cycle PCR from bulk P1 cerebellum. $N = 1{,}535$, 1,258, 525, 539, 248, 189, 959, 455, 83 and 108 for the ten bars from left to right. Vertical gray lines indicate 95% confidence intervals. (**b**) Number of isoforms added to the annotation and number of affected genes. (**c**) Percentage of complete unique isoforms from rep2 that could also be observed in rep1 (in any cell type) broken up by cell type of origin from rep1. $N = 1{,}059$, 1,350, 356, 277, 546, 337, 969, 347, 456, 105 and 216 for the 11 bars from left to right. Vertical gray lines indicate 95% confidence intervals. (**d**) Percentage of complete unique replicated isoforms from rep2 that could also be observed in rep1 (in the same cell type) broken up by cell type of origin from rep1. $N = 807$, 656, 594, 208, 128 and 56 for the six bars from left to right. Vertical gray lines indicate 95% confidence intervals. (**e**) Distributions of coverage with microglial short reads for introns in the enhanced annotation that were exclusively observed in one cell type (indicated by name under the *x* axis). Boxplots elements (quartiles, median, whiskers) are standard elements as defined in the 'boxplot' function of *R*. (**f**) Single-gene view for the *Bin1* gene, the second most Alzheimer's-disease-associated gene. Bottom, black track: GENCODE annotation. Blue track, ScISOr-Seq data with each line representing one molecule for neuronal cells. Red-brown track, ScISOr-Seq data with each line representing one molecule for astrocytes. Purple track, ScISOr-Seq data with each line representing one molecule for microglia. Gray track, ScISOr-Seq data with each line representing one molecule for OPCs. Orange exons indicate alternative internal exons used in at least three molecules as well as novel internal exons.

ScISOr-Seq generates chimeric transcripts, we obtained 164 million 150-bp paired-end reads on bulk RNA from P1 cerebella using a six-cycle short-range PCR after RNA fragmentation. Based on this experiment, we confirmed 91.6–97.6% of the novel ScISOr-Seq junctions across different cell types (Online Methods and **Fig. 2a**). To reduce the effect of PCR artifacts on the improved mouse Gencode annotation to a minimum, and to allow for adding transcripts expressed at low levels, we produced an enhanced cell-type-resolved annotation that had good six-cycle PCR short-read support. For each added isoform, each intron and internal exon was required to be annotated in Gencode, or to be supported by two or more six-cycle PCR short reads (Online Methods), resulting in 16,872 isoforms for 6,927 genes (**Fig. 2b** and **Supplementary Data Set 1**). The barcode attached to each of these isoforms indicates the single cell and the cell type of origin of each isoform. 42.8% (7,219 of 16,872) of the added isoforms had at least one splice site not annotated in Gencode. With respect to the UCSC[34] and RefSeq[35] annotations, 94.0 and 70.9%, respectively, of added isoforms were novel (Online Methods and **Supplementary Data Set 2**). We performed ScISOr-Seq for rep2 (PB_rep2), albeit at an approximately fourfold lower sequencing depth (1.3 million CCS, compared with 5.2 million for PB_rep1). 65.7 (microglia) to 76.2% (neuronal progenitors) of new PB_rep2 isoforms were also present as PB_rep1 isoforms (irrespective of the cell type they were assigned to in rep1; **Fig. 2c**). Given replication of an isoform in any cell type, cell-type-specific replication of a PB_rep2 isoform in the same cell type in rep1 reached 70–80% in larger clusters, but only reached lower percentages in smaller clusters with substantially fewer long reads (**Fig. 2d**). To validate the correct calling of the individual cell of origin for each isoform, we performed immunopanning to specifically isolate microglia in P1 cerebella followed by short-read RNAseq (P1_CB_Microglia). P1_CB_Microglia reads were compared with all isoforms originating from a single microglial cell (and then with isoforms of single cells belonging to other cell types). This confirmed the microglial origin of long-read junctions exclusively observed in microglial single-cell long reads as compared with junctions observed exclusively in non-microglial single-cell long reads

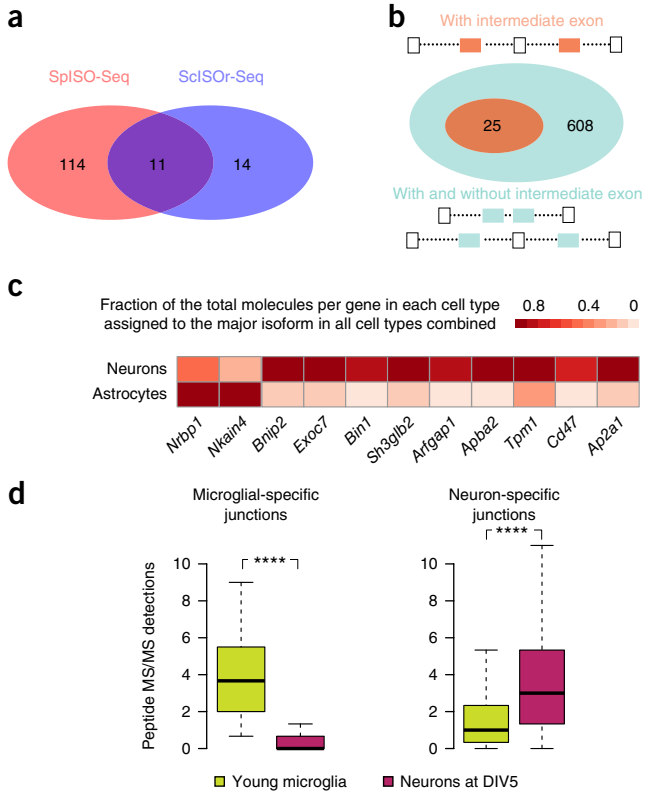

Figure 3 Quantitative isoform analysis. (**a**) Venn diagram of genes with exon coordination (with intermediate exons) between mouse ScISOr-Seq data and human SpISO-Seq data. (**b**) Venn diagram of genes with exon coordination requiring the presence of intermediate exons and not requiring it. (**c**) Relative (to total number of molecules from the gene in the cell type in question) isoform abundance of the major (in all cell types combined) isoform of genes having significantly different major isoform abundance between neurons and astrocytes. (**d**) Microglial and neuronal peptide coverage from label-free proteomics for ScISOr-Seq junctions repeatedly observed in microglia (but not in neurons) (left) and for ScISOr-Seq junctions repeatedly observed in neurons (but not in microglia) (right). ****P values of $<2.2 \times 10^{-16}$. Boxplots elements (quartiles, median, whiskers) are standard elements as defined in the 'boxplot' function of *R*.

(**Fig. 2e**). Similarly, immunopanning for astrocytes and BG (both marked by *Hepacam*), and OPCs (which are known to be enriched in *Hepacam*-sorting) and short-read sequencing (P1_CB_Astrocytes) revealed that the highest coverage for junctions was exclusively in astrocyte, BG and OPC ScISOr-Seq isoforms (**Supplementary Fig. 7a**). This was more pronounced for junctions observed three or more times in ScISOr-Seq data in one cell type (**Supplementary Fig. 7b**). These immunopanning data indicate that junctions observed only in astrocytes, BG and OPCs are also expressed at a lower level in other cell types originating from the same stem cell.

We next examined alternative splicing in the *Bin1* gene, which is important in Alzheimer's disease and is expressed in multiple cell types[36]. In addition to four annotated alternate exons (A1, A3, A4 and A5; **Fig. 2f**) in mouse Gencode for *Bin1*, we found two more alternate exons, A2 and A6, in ≥3 reads. These two exons, located before and after the three-alternate-exon block A3–5, are also alternate in human Gencode (version 24) annotation[32], in our human brain isoform sequencing[7] (Figure S5 in ref. 6) and in previous publications[37–39] (referred to as 12a and 13 (refs. 37,39) and 13 and 17 (ref. 38)). Notably, the major isoform of single cells of neuronal subtypes (inclusion of A1–4 and A6, but not A5) and the major isoform of single cells of microglia and oligodendrocytes (skipping of exons A1–6) are not annotated in mouse Gencode. The combined six alternate exons A1–6 (four known and two novel in mouse) of *Bin1* allow for $64 = 2^6$ combinations, of which Gencode includes two, whereas our data reveal 12 additional ones. Of the 12 added isoforms, 5 isoforms were supported by 39, 16, 6, 3 and 2 UMIs, and seven novel isoforms were supported by one UMI each (**Fig. 2f**).

Skipping of *Bin1* alternative exons A1 and A2–6 (entirely or partially observed in human[6,7,37–39]) occurred in all of the microglial reads and in most of the astrocyte and oligodendrocyte reads, but not in neuronal subtypes (two-sided Fisher test P value = 1.8×10^{-8} neurons versus microglia). This indicates that the coordination of distant alternate exons in *Bin1*, which is also observed in adult human brain[6,7], is a result of cell-type-specific isoform expression of the isoforms using all or none of exons A1–6. Coordination of alternate exons is of crucial biological importance[6,7,40], so we searched for this in our ScISOr-Seq data. We found 25 genes with coordination of alternate exons that were separated by intermediate exons. These genes overlapped with coordinated human genes published previously[7]. These coordination events were therefore observed using different methods (ScISOr-Seq versus SpISO-Seq), in different species (mouse versus human), different age samples (mouse P1 versus human adults) and in different tissues (cerebellum versus entire brain) (**Fig. 3a**). Testing all exon pairs, adjacent or separated by intermediate exons, we found 633 genes with coordination, including all 25 with intermediate exons. Thus, most coordinated pairs were adjacent exon pairs (**Fig. 3b**). To explore the underlying causes of coordination, we tested complex genes (Online Methods) for differences in relative major isoform abundance between the groups of neuronal single cells and single astrocytes. Using *de novo* junctions increased the confidence of detection of differential splicing[41,42], and our procedure took into account known and novel splice sites. We observed such changes in relative abundance in 11 of these genes, and of these *Bin1*, *Cam2g*, *Exoc7*, *Nkain4* and *Zdhhc20* were also coordinated. 20% (5 of 25) of coordination events of alternate exons, which were separated by constitutive exons, were a result of differences in isoform abundance between neurons and astrocytes (**Fig. 3c**). However, for adjacent alternative exon pairs, only 3.5% (21 of 608, two-sided Fisher P value < 0.005) were a result of cell-type-specific isoform abundance between neurons and astrocytes. Neuronal and microglial peptides have recently been identified in young mouse brains[43]. To further validate quantitation with ScISOr-Seq, we analyzed junctions that were observed repeatedly (Online Methods) in microglia, but not in other neuronal subtypes. These junctions revealed much higher coverage (fold change of microglial to neuronal mean: 11.1) with microglial peptides than with neuronal peptides (**Fig. 3d**). Conversely, junctions that were observed in neurons, but not in microglia, had higher peptide coverage (fold change of neuronal to microglial mean: 2.3) in neuronal mass-spectral data sets[43] (**Fig. 3d**). Thus, notwithstanding the higher PCR-cycle number and lower abundance of reads using ScISOr-Seq rather than SpISO-Seq, ScISOr-Seq can be applied in quantitative analyses.

In the future, we envisage that for genetic risk factors for brain disorders, for example, Alzheimer's-disease-related genes, such as *MAPT*, *BIN1* and *APOE*, the effects of disease-associated single-nucleotide polymorphisms can be explained by analyzing cell-type-specific isoform expression using ScISOr-Seq. Here, we investigated a cerebellum at postnatal day 1; it will be interesting going forward to evaluate splicing alterations between P1 and adult cerebellum as

cells mature to form adult cerebellum. Our full-length RNAs from single cells cover all of the single-nucleotide polymorphisms in the coding region of mature RNA and may help to attribute single cells to a specific individual[44] in pooled samples.

ScISOr-Seq has limitations. Multiple deeply sequenced replicates are needed for precise quantification. This is inexpensive using short-read sequencing, so precise statistics for quantification in bulk RNA-Seq are readily obtained[41,42]. Use of long-read technology in ScISOr-Seq makes accurate quantification expensive for now. Our estimates for specificity and sensitivity of barcode recognition in long reads are based on using 16-mer 10xGenomics barcodes for 6,000–7,000 cells. Ideally, users would perform analyses of pairwise barcode distances after every 10xGenomics short-read run. If the number of cells is increased to >1 million while still relying on 16-mer barcodes, we would advise reassessment of specificity and sensitivity, as specificity is likely to drop. Likewise, using shorter barcodes might reduce specificity. Generally, larger cell numbers will require longer barcodes.

In summary, ScISOr-Seq enables long-read full-length RNA-Seq in single cells that can be clustered into cell types with a very low identification error rate.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This work used the Genomics Resources Core Facility and owes special thanks to J. Xiang and A. Wan. This work was supported by start-up funds (Weill Cornell Medicine) and a Leon Levy Fellowship in Neuroscience to H.U.T. as well as an R01 from the National Institute of Neurological Disorders and Stroke (1R01NS105477) to M.E.R.

AUTHOR CONTRIBUTIONS

P.G.C., I.G., S.A.S. and H.U.T. devised the experiments. P.G.C., B.H., I.G., S.A.S., O.F. and W.L. performed the experiments. I.G., A.B.S. and H.U.T. devised the analyses. I.G., A.M., A.J., T.F., F.K. and H.U.T. performed the analyses. All of the authors discussed and interpreted the results throughout the project. I.G. and H.U.T. wrote the paper with inputs from all of the other authors. B.B., M.E.R. and H.U.T. supervised the project.

COMPETING INTERESTS

The authors declare no competing interests.

1. Sharon, D., Tilgner, H., Grubert, F. & Snyder, M. A single-molecule long-read survey of the human transcriptome. *Nat. Biotechnol.* **31**, 1009–1014 (2013).
2. Au, K.F. *et al.* Characterization of the human ESC transcriptome by hybrid sequencing. *Proc. Natl. Acad. Sci. USA* **110**, E4821–E4830 (2013).
3. Oikonomopoulos, S., Wang, Y.C., Djambazian, H., Badescu, D. & Ragoussis, J. Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations. *Sci. Rep.* **6**, 31602 (2016).
4. Tilgner, H., Grubert, F., Sharon, D. & Snyder, M.P. Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proc. Natl. Acad. Sci. USA* **111**, 9869–9874 (2014).
5. Koren, S. *et al.* Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. *Nat. Biotechnol.* **30**, 693–700 (2012).
6. 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).
7. Tilgner, H. *et al.* Microfluidic isoform sequencing shows widespread splicing coordination in the human transcriptome. *Genome Res.* **28**, 231–242 (2018).
8. Bolisetty, M.T., Rajadinakaran, G. & Graveley, B.R. Determining exon connectivity in complex mRNAs by nanopore sequencing. *Genome Biol.* **16**, 204 (2015).
9. Roy, C.K., Olson, S., Graveley, B.R., Zamore, P.D. & Moore, M.J. Assessing long-distance RNA sequence connectivity via RNA-templated DNA–DNA ligation. *eLife* **4**, e03700 (2015).
10. Treutlein, B., Gokce, O., Quake, S.R. & Südhof, T.C. Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing. *Proc. Natl. Acad. Sci. USA* **111**, E1291–E1299 (2014).
11. Schreiner, D. *et al.* Targeted combinatorial alternative splicing generates brain region–specific repertoires of neurexins. *Neuron* **84**, 386–398 (2014).
12. Karlsson, K. & Linnarsson, S. Single-cell mRNA isoform diversity in the mouse brain. *BMC Genomics* **18**, 126 (2017).
13. Byrne, A. *et al.* Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat. Commun.* **8**, 16027 (2017).
14. Song, Y. *et al.* Single-cell alternative splicing analysis with expedition reveals splicing dynamics during neuron differentiation. *Mol. Cell* **67**, 148–161 (2017).
15. Shalek, A.K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236–240 (2013).
16. Zeisel, A. *et al.* Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138–1142 (2015).
17. Lake, B.B. *et al.* Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science* **352**, 1586–1590 (2016).
18. Zheng, G.X.Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
19. Darmanis, S. *et al.* A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci. USA* **112**, 7285–7290 (2015).
20. Jaitin, D.A. *et al.* Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 (2014).
21. Pollen, A.A. *et al.* Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* **32**, 1053–1058 (2014).
22. Fededa, J.P. *et al.* A polar mechanism coordinates different regions of alternative splicing within a single gene. *Mol. Cell* **19**, 393–404 (2005).
23. Fagnani, M. *et al.* Functional coordination of alternative splicing in the mammalian central nervous system. *Genome Biol.* **8**, R108 (2007).
24. Mecklenburg, N. *et al.* Growth and differentiation factor 10 (Gdf10) is involved in Bergmann glial cell development under Shh regulation. *Glia* **62**, 1713–1723 (2014).
25. Koirala, S. & Corfas, G. Identification of novel glial genes by single-cell transcriptional profiling of Bergmann glial cells from mouse cerebellum. *PLoS One* **5**, e9198 (2010).
26. Lein, E.S. *et al.* Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**, 168–176 (2007).
27. Butts, T., Green, M.J. & Wingate, R.J.T. Development of the cerebellum: simple steps to make a 'little brain'. *Development* **141**, 4031–4041 (2014).
28. Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).
29. Waterston, R.H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
30. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
31. Tilgner, H. *et al.* Accurate identification and analysis of human mRNA isoforms using deep long read sequencing. *G3* **3**, 387–397 (2013).
32. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
33. Mcmanus *et al.* Global analysis of trans-splicing in *Drosophila*. *Proc. Natl. Acad. Sci. USA* **107**, 12975–12979 (2010).
34. Karolchik, D. *et al.* The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.* **42**, D764–D770 (2014).
35. O'Leary, N.A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
36. Zhang, Y. *et al.* Purification and characterization of progenitor and mature human astrocytes reveals transcriptional and functional differences with mouse. *Neuron* **89**, 37–53 (2016).
37. Ge, K. *et al.* Mechanism for elimination of a tumor suppressor: aberrant splicing of a brain-specific exon causes loss of function of Bin1 in melanoma. *Proc. Natl. Acad. Sci. USA* **96**, 9689–9694 (1999).
38. Fugier, C. *et al.* Misregulated alternative splicing of BIN1 is associated with T tubule alterations and muscle weakness in myotonic dystrophy. *Nat. Med.* **17**, 720–725 (2011).
39. Karni, R. *et al.* The gene encoding the splicing factor SF2/ASF is a proto-oncogene. *Nat. Struct. Mol. Biol.* **14**, 185–193 (2007).
40. Anvar, S.Y. *et al.* Full-length mRNA sequencing uncovers a widespread coupling between transcription initiation and mRNA processing. *Genome Biol.* **19**, 46 (2018).
41. Vaquero-Garcia, J. *et al.* A new view of transcriptome complexity and regulation through the lens of local splicing variations. *eLife* **5**, e11752 (2016).
42. Li, Y.I. *et al.* Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* **50**, 151–158 (2018).
43. Sharma, K. *et al.* Cell type– and brain region–resolved mouse brain proteome. *Nat. Neurosci.* **18**, 1819–1831 (2015).
44. Kang, H.M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).

ONLINE METHODS

Ethics statement. All experiments were conducted in accordance with relevant NIH guidelines and regulations, related to the Care and Use of Laboratory Animals tissue. Animal procedures were performed according to protocols approved by the Research Animal Resource Center at Weill Cornell College of Medicine.

Animals and tissue isolation. C57BL/6 mice purchased from Taconic Biosciences were maintained following the approved protocol. For each replicate, the cerebellum was dissected from a single P1 neonatal mouse.

Tissue disassociation. Dissected cerebellum was placed into 2.5 ml Hibernate E/B27/GlutaMax medium (BrainBits cat#HEB). HEB medium was removed and kept for later trituration steps. Tissue was incubated with 2 ml of 2 mg/ml activated papain (BrainBits cat#PAP) for 25 min at 37 °C with gentle mixing. After allowing tissue to settle, papain was removed and replaced with the retained HEB medium and tissue was gently triturated 15–20 times. The debris was allowed to settle and the supernatant was centrifuged at 400 rcf for 2 min. The cell pellet was re-suspended in 500 µl neuronal culture medium NbActiv1 (BrainBits cat#Nbactiv1), filtered through a 30-µm cell strainer (Miltenyi Biotec cat#130-041-407) and were diluted to 1,000 cells/µl in NbActiv1 for capture on the 10x Genomics Chromium controller.

10x Genomics single-cell capture. The disassociated cells were captured on the 10x Genomics Chromium controller according to the Chromium Single Cell 3′ Reagent Kits V2 User Guide (10x Genomics PN-120237) with the following modification. PCR cycles were increased, from the recommended ten cycles for recovery of 8,000 cells, to 16 cycles to target a yield of cDNA enabling simultaneous Illumina and PacBio library preparation.

Illumina and Pacific Biosystems library preparation. Illumina library preparation was performed using 100 ng of amplified cDNA following the Chromium Single Cell 3′ Reagent Kits V2 User Guide (10x Genomics PN-120237) reducing final indexing PCR cycles to ten cycles from the recommended 14 cycles to increase library complexity. Sequencing was performed on the Illumina NextSeq 500 (Illumina cat#FC-404-2001) with a 26-bp (read one) and 98-bp (read two) run mode. PacBio library preparation was performed with 850 ng of amplified cDNA using SMRTbell Template Prep Kit (PacBio cat# 100-991-900) to obtain Sequel compatible library complex to be sequenced using appropriate number of SMRTcells (PacBio cat#101-008-000).

Total cerebellum short-read Illumina library preparation using six cycles of PCR (and no full-length PCR). Illumina compatible libraries were produced from 700 ng total RNA using NEBNext Ultra II RNA Library Prep Kit (NEB Cat#E7770S) following manufactures protocol with the following modifications. Target insert size was 450 bp for compatibility with paired end 150-bp sequencing mode. Number of PCR cycles was reduced to 6 to limit the effect of PCR aberrations on the final library. Sequencing was performed on the Illumina NextSeq 500 instrument.

minION library preparation and sequencing. Library was prepared using 1ug cDNA from 10x Genomics following the 1D[2] Sequencing of Genomic DNA (SQK-LSK308) protocol by Oxford Nanopore and sequenced using MinIon FLO-MIN107 flowcell over 48 h. Base calling was performed using Albacore version 2.1.3 to obtain fastq files.

Pacific Biosystems barcoded Iso-Seq. 100ng of total RNA from GM12878 cell line was used as input into the Smart-Seq2 protocol[45] with the following modifications. A custom oligo(dT) primer containing the 25-bp Smart-Seq Primer Site, 21-bp 10x adaptor, 16-bp cell barcode and 5-bp UMI was used for reverse transcription.

Custom primer used: AAGCAGTGGTATCAACGCAGAGTACCTACACG ACGCTCTTCCGATCGATCAGTTCACGCATANNNNNTTTTT TTTTTT TTTTTTTTTTTTTTTTTTTTVN.

To mimic 10x Genomics cDNA amplification, 16 cycles of PCR were performed and PacBio sequencing was carried out with 1 µg of cDNA as described above.

Alignment, mapping and data analysis of Illumina short reads. The 10x cellranger pipeline (version 2.0.0) was run on the raw Illumina sequencing data to obtain a single-cell expression matrix object. The command lines were as follows.

```
cellranger mkfastq –id = scISOr-Seq –run = path_to_
raw_data –csv = path_to_SampleSheet.csv

cellranger count –id = scISOr-Seq –transcriptome =
path_two_refdata-cellranger-mm10-1.2.0 –fastqs =
/fastqpath/–sample = scISOr-Seq
```

The resulting matrix was read into *R* using the *Seurat* package[46] (version 2.2.1). For replicate 1, cells that had unique gene counts over 2,500 (duplets) or less than 200 (background) were removed from further analysis. For replicate 2, with higher sequencing depth and more cell death (high mitochondrial gene expression), we removed cells that had over 2,700 genes (duplets) or less than 200 genes (background) or greater than 0.15% mitochondrial gene expression. Next steps were same for both the replicates. The number of UMIs, percentage of mitochondrial gene expression were regressed from each cell and then the gene expression matrix was 'logNormalized' and scaled to 10,000 reads per cell. Next, we clustered the cells using 20 principal components (PCs) using Shared-Nearest-Neighbor (SNN) algorithm with a 0.6 resolution. Each cluster was annotated with a cell-type based on the expression of known cell-type specific marker genes as described in the main text (**Fig. 1b**). The set of marker genes for each annotated cluster was calculated using Wilcoxon rank sum test to obtain genes that were expressed in at least 25% of the cells of a cluster with a 0.25 log fold change gene expression higher than all other clusters. The reproducibility of clusters was tested using the Jaccard index *J* (**Supplementary Fig. 3a–c**) between two clusters calculated as

$$J(X_1, X_2) = \frac{|X1 \cap X2|}{|X1 \cup X2|}$$

where X_i is the set of marker genes for cluster *i*. **Supplementary Code** contains an R markdown detailing the above analysis.

Generation of circular consensus reads. Using the default SMRT-Link (version 5.0.1.9585) parameters, we performed CCS as follows with the following modified parameters: maximum subread length 14,000, minimum subread length 10 and minimum number of passes 3.

Poly(A) tail and barcode detection in long reads. Barcode assignment and read filtering were performed in the three following steps, polyA tail detection (finding a continuous stretch of 9 alanines within 200 bp of either read end assuming 2/30 error rate in homopolymer sequencing), barcode matching (reads searched for perfect match to 16-bp 10× cell-type associated barcodes for example 6,627 barcodes in rep1, along with matching to all the 6,627 barcodes containing 1-bp substitution and indel error), and chimera assignment (reads with two polyA tails with greater than 90% alanines). We use the reads with only one distinct polyA tail, with a unique match to one of the 10× cell-type assigned barcodes only and with no matches with 10× cell-type assigned barcodes with 1 error.

Library complexity estimation and probability of a unique molecule being sequenced twice. Let us denote for an individual cDNA molecule the probability to be detected with long-read sequencing as $P = \mathrm{P}(N \geq 1)$ and the probability to be sequenced at least twice given at least one observation as $\mathrm{P}(N \geq 2 | N \geq 1)$. We have previously given[6] a mathematical proof (which is more than likely to be in many textbooks as well, due to its basic nature in probability theory) for the following upper bound on $\mathrm{P}(N \geq 2 | N \geq 1)$:

$$\frac{p + \ln(1-p) * (1-p)}{p}$$

We observe a median of 260 distinct UMIs per cell in the long read data and a median of 3,875 UMIs in the short read data. The assumption that the short read data has picked up all molecules yields an upper bound on the detection probability $\mathrm{P}(N \geq 1)$ as 260/3,875, giving 6.7%.

Using formula 1, we can then calculate $P(N \geq 2|N \geq 1)$ as 0.034 or 3.4%. In practice, we observe a $P(N \geq 2|N \geq 1)$ of 3.84% (10 out of 260; a median of 260 detected UMIs per cell for a median of 270 long reads per cell), which is roughly consistent with the theoretical prediction.

Simulation of barcode identification correctness. We used 42 million R1 reads from rep1, to construct 77-mers containing 10× sequencing adaptor (21 bp), the single cell barcode from the short reads (16 bases), the UMI (10 bases) along with 30 threonines of a theoretical polyA-tail. Given that we used real 10× data from rep1 it included reads containing barcodes not assigned to clustered cell during Seurat analysis (only 73.9% of the reads contain a barcode assigned to a single cell). We employed an error model with 0.1% probability for substitutions (three distinct substitutions being possible at every nucleotide), 0.25% deletion probability (one deletion being possible at every nucleotide) and 0.25% insertion probability (four insertions being possible after every nucleotide). This approximately represents an error probability of 1.5%. We used this error model to insert PacBio errors in silico into these 77-mers and performed polyA-tail and barcode detection. On average (across repetitions) in 41.3% of the 77-mers, we discard the read, because we do not find any of the 6,627 known cell barcodes. This defines the sensitivity of 58.7% to the procedure (compared to 73.9% for Illumina 3′seq). Non-perfect sensitivity increases the cost of the experiment but in no way causes any false positive barcode identification. For the reads, for which we do assign a barcode, we find (on average across repetitions) that in 99.99% of the 77mers we assign the correct single cell, leading to an estimate of specificity of 99.99%.

Alignment of long reads. In our previous publications, we have used GMAP[47] to align long reads to the genome, which for synthetic long reads[6,7] and PacBio reads[1,4] gave highly satisfactory results. In the earlier publications using PacBio data, we determined the mapping quality ourselves, counting mismatches, insertions and deletions. Starting in 2015 (ref. 6), we have relied on the MAPQ field in sam files to determine trustworthy alignments (see above). We found that GMAP gave very low MAPQ scores, which led to all reads being filtered. We did not observe this with STARlong and therefore chose STARlong as the aligner. We aligned long reads to the mouse genome[31] (version mm10), using the STARlong aligner[30] with the following parameters (recommended by Pacific Biosciences):--outSAMattributes NH HI NM MD, --readNameSeparator space, --outFilterMultimapScoreRange 1, --outFilterMismatchNmax 2000, --scoreGapNoncan -20, --scoreGapGCAG -4, --scoreGapATAC -8, --scoreDelOpen -1, --scoreDelBase -1, --scoreInsOpen -1, --scoreInsBase -1, --alignEndsType Local, --seedSearchStartLmax 50, --seedPerReadNmax 100000, --seedPerWindowNmax 1000, --alignTranscriptsPerReadNmax 100000 and --alignTranscriptsPerWindowNmax 10000.

Analysis of long-read mappings is described in **Supplementary Note 1**.

An enhanced annotation using only novel isoforms, for which all splice sites are known. Using the novelty definition as described previously[31] (**Supplementary Note 1**), we determined all long reads that (1) had been assigned to a named cell cluster; (2) employed only splice sites that existed already in the Gencode vM10 annotation; (3) shared more splice sites with one gene than with all other genes (so that a clear-cut gene assignment for the long-read mapping could be made); and (4) for which each splice junction was observed at least twice in the combined ScISOr-Seq data set.

Relying only on long-reads, for which all splice sites are known in the annotation, is meant to increase specificity of the produced annotation (because PacBio sequencing and mapping artifacts can lead to wrong splice site assignments) at the cost of missing some true positive novel isoforms.

Each long read defines an ordered list of introns $X = (x_1, x_2,\ldots, x_n)$.

- If this ordered list of introns is unique among all long-reads, the long-read is added as is to the annotation, with its hypothetical TSS being defined as its first mapped nucleotide and its hypothetical polyA-site being defined as its last mapped nucleotide. The cell type, from which the long read originates, is then listed as the origin of the isoform in the enhanced annotation.
- If multiple novel read alignments (with respect to the Gencode vM10 annotation) share the same ordered list of introns $X = (x_1, x_2,\ldots, x_n)$, the hypothetical TSS is defined as the most upstream mapped nucleotide across all these reads. Likewise the hypothetical polyA-site is defined as the most downstream mapped nucleotide across all these reads. All the cell types, from which these long reads originate are given as the origin(s) of the isoform in the enhanced annotation.

An enhanced annotation using also novel isoforms with novel splice sites. In order to produce a second, more sensitive and more reproducible annotation, we used only novel isoforms, for which:

- Each junction was either annotated or observed at least twice in short-read sequencing of bulk P1 cerebellum using only six PCR cycles on the library after RNA fragmentation.
- Each internal exon was annotated or observed at least twice in the same six-cycle PCR short-read sequencing experiment
- Each introns was shorter than 77,048 bases (the 99th percent quantile of annotated Gencode introns)

Confirmation of novel junctions with a six-cycle PCR short-read RNA-sequencing experiment (PCR executed on the library after RNA fragmentation). Experimental procedures for the six-cycle PCR are described above. We aligned the resulting 150-bp paired-end reads to the mouse genome[31] (version mm10) using STAR[30] and the following parameters:--readFilesCommand zcat, --outFilterMultimapNmax 1, --outFilterIntronMotifs, RemoveNoncanonical, --outFilterMismatchNmax 5, --alignSJDBoverhangMin 6, --alignSJoverhangMin 6, --outFilterType BySJout, --alignIntronMin 25, --alignIntronMax 1000000, --outSAMstrandField intronMotif, --outSAMunmapped Within, --runThreadN 24, --outStd SAM, and --alignMatesGapMax 1000000.

STAR reports high confidence junctions. A ScISOr-Seq junction was considered validated, if it was reported by STAR using the 6-cycle PCR approach.

Novelty of isoforms with respect to the UCSC and RefSeq annotations. We downloaded the UCSC[34] and RefSeq[35] annotations from the UCSC table browser. Novelty of isoforms in our enhanced annotation was checked against these annotations exactly as outlined for the Gencode annotation (**Supplementary Note 1**).

Analysis of cell-type-specific isoform abundance. For a given gene, we determined all internal exons (known or novel) that were present in at least 5% of all molecules and at most 95% of all molecules. A gene is considered a complex gene if it contains at least two alternative exons that are separated by one or more constitutive exon(s).

For complex genes, we produced a Boolean matrix of reads times alternative exons. The exon combination being most frequent in this matrix was called the major isoform. We then separated the matrix into one matrix per cell type. For the major isoform, we determined the number of reads supporting it and not supporting it in one cell type and then separately in another cell type. These four numbers were then used to populate a 2×2 contingency table and a two-sided Fisher test was applied. We corrected for multiple testing (over all genes) using the Benjamini-Hochberg[48] method.

Generating translated exon-exon junction crossing theoretical peptides. For each unique junction in our (cell-type-assigned) ScISOr-Seq data, we retrieved 30 upstream and 30 downstream nucleotides. These 60 nucleotides were translated in 6 frames.

Junctions exclusively observed in neurons or microglia and comparison against peptides. ScISOr-Seq junctions observed at least 6 times in neurons and never in microglia were termed 'neuron-specific junctions' and ScISOr-Seq junctions observed at least three times in microglia and never in neurons were termed 'microglia-specific junctions'.

Matching of hypothetical translated sequences against published mass-spec data. We compared microglial-specific and neuron-specific junctions from the ScISOr-Seq data to a previously published proteomic analysis[43], publicly

doi:10.1038/nbt.4259

available at ProteomeXchange with data set identifier PXD001250. We used the MaxQuant analysis results provided in 'search_v1_SingleShot_And_Library_Matched_Within.zip'. From the provided 'peptides.txt' table containing identified and quantified peptides, we removed reverse identifications from the decoy database, peptides matching the included contaminant database and identifications with posterior error probability larger than 5%.

In silico tryptic digestion of sequences from the ScISOr-Seq data enabled matching to the tryptic peptides identified in proteomics. First, input FASTA sequences were truncated at stop codons. After *in silico* tryptic digestion with at most 1 missed cleavage allowed we removed tryptic peptides that match the begin or end (unless there is a trailing stop codon) of the (arbitrarily truncated) FASTA input sequences. Finally, we map between proteomic and ScISOr-Seq data by exact string matching of tryptic peptide sequences.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability. The raw sequencing data can be accessed from the Sequence Read Archive under the bioproject PRJNA428979 with the following accession numbers and associated data set names: SRR7345562 (PB_rep1), SRR7652917 (PB_rep2), SRR7345560 (1d), SRR7345559 (1d2.pass), SRR7345558 (1d2.fail), SRR7623730 (P1_CB_lowCycle), SRR7617314 (P1_CB_Microglia), SRR7617315 (P1_CB_Astrocytes). The processed data for Single cell and cluster level isoform expression can be queried at isoformAtlas.com.

45. Picelli, S. *et al*. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
46. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
47. Wu, T.D. & Watanabe, C.K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
48. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).

nature research

Corresponding author(s): Hagen U Tilgner

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Python, Bash, 10x cellranger pipeline (version 2.0.0), Albacore version 2.1.3
Data analysis	Bash, Awk, Python, R, STAR, STARlong

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The raw sequencing data can be accessed from the Sequence Read Archive under the bioproject PRJNA428979 with the following accession numbers and associated dataset names: SRR7345562 (PB_rep1), SRR7652917 (PB_rep2), SRR7345560 (1d), SRR7345559 (1d2.pass), SRR7345558 (1d2.fail), SRR7623730

Nature Biotechnology: doi:10.1038/nbt.4259

(P1_CB_lowCycle), SRR7617314 (P1_CB_Microglia), SRR7617315 (P1_CB_Astrocytes). The processed data for Single cell and cluster level isoform expression can be queried at isoformAtlas.com. The mass spec data used in this study is available from www.mousebrainproteome.com.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was determined by the microfluidic capture efficiency of the single cells and the fidelity of gene expression counts within each cell. We targeted for capturing 6000-8000 single cells in order to have at least 30 single cells to perform statistics in the smallest cell cluster representing 1% of the total captured cells. We obtained 6627 single cells from replicate 1, 6506 single cells from replicate 2 from a P1 mouse cerebellum. And therefore, could capture microglial cells that are present about 1% of all the cells from the cerebellum.
Data exclusions	No data was excluded
Replication	Data from single cell Isoform sequencing experiments were replicated in two biological replicates to obtain reproducible data from 11 cell-types. We also performed orthogonal experiments to replicate data. We confirmed more than 90% of novel junctions described by single cell Isoform sequencing across all the 11 cell types by low PCR deep RNA-Seq from whole P1 cerebellum. We also confirmed cell-type specific novel junctions by deep RNA-Seq from isolated Microglia and Astrocytes from P1 cerebellum using immunopanning. We also confirmed cell-type specific junctions by mining publicly available Mass-Spec data.
Randomization	The mice used in the study were not random. All the mice used in the study were male to control for sex as a covariate. For the immunopanning experiments, inter individual differences were controlled for by pooling 5 or more mice brain from the same litter at a time.
Blinding	Blinding not relevant to our study because it is not a subjective trial and the results presented here are purely based on objective description of our novel experimental technology.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	Rat anti-mouse CD45 (BD Pharmingen, 550539), Mouse anti HepaCAM (R&D systems, MAB4108). Final dilution on panning plate for CD45: 0.1ug/ml , HepaCAM 0.5 ug/ml. The lot number for the antibodies used in the study were unavailable. However, the antibodies are monoclonal with little variation between lots as evidenced by their extensive and reliable use in the community for many years.
Validation	For validation for CD45 and HepaCAM, see Zhang, Sloan et al. 2016 (Neuron), Zhang, Sloan et al. 2014 (J Neuro), Foo et al. 2011 (Neuron)

Nature Biotechnology: doi:10.1038/nbt.4259

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	C57BL/6 male mice were purchased from the Taconic Biosciences. The mice used in the study were from developmental time point P1 (1 day after birth)
Wild animals	Study did not involve wild animals
Field-collected samples	Study did not involve field collected samples

Nature Biotechnology: doi:10.1038/nbt.4259