Molecular Tools for Transcriptomic Measurements

Charles M Roco

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:
Dr. Georg Seelig, Chair
Dr. James Carothers
Dr. Barry Lutz

Program Authorized to Offer Degree:
Bioengineering

University of Washington

Abstract

Molecular Tools for Transcriptomic Measurements

Charles M Roco

Chair of the Supervisory Committee:
Associate Professor Georg Seelig
Electrical Engineering and Computer Science & Engineering

Despite tremendous advances in next generation sequencing, we are still far from understanding our own genome. Much of the sequencing we do today measures genotype, providing valuable insights to the makeup of our DNA. However, to really understand our genome we must get closer to phenotype. We need methods that capture what biological functions are being carried out in a tissue. Transcriptome sequencing provides a snapshot of gene expression, providing insights on what proteins and subsequent function might be present.

This thesis discusses the development of new technologies that build on the current tools for transcriptome level measurements. First I will discuss the development of a single-cell RNA-seq method we call SPLiT-seq, which labels the cellular origin of RNA through combinatorial barcoding. SPLiT-seq is compatible with fixed cells or nuclei, allows efficient sample multiplexing and requires no customized equipment. To demonstrate the power of SPLiT-seq, we analyzed 156,049 single-nucleus transcriptomes from postnatal day 2 and 11 mouse brains and spinal cords. Over 100 cell types were identified, with gene expression patterns corresponding to cellular function, regional specificity, and stage of differentiation. Pseudotime analysis revealed transcriptional programs driving four developmental lineages, providing a snapshot of early

postnatal development in the murine central nervous system. Next, I will describe a targeted enrichment method we call CleavR to increase detection of rare molecules while bringing sequencing costs down. CleavR selectively captures molecules comprising a user-defined sequence by leveraging the highly specific and active nature of rnases. Finally, I will discuss our development of a massively parallel reporter splicing assay to measure how DNA mutations impact alternative splicing.

It is my hope that this thesis will serve a resource for the entire genomics community - whether scientific experts adopt one of these technologies to gain further understanding of their work, or other technology enthusiasts use it to build better tools for the future.

Acknowledgments

I could not have asked for a better experience during my time in this PhD program. The friends I've met and the science I've learned along the way is something that I will always cherish. I am indebted to the many people that have helped me get here, both before and during my time at the University of Washington.

Georg Seelig has been a fantastic mentor. He gave me the freedom to develop my own experiments and try out new ideas, even if they were out of the lab's comfort zone. At the same time, this more hands-off approach was paralleled by always being there when it was needed, providing invaluable advice in both science and career related choices.

A large portion of my time spent during this program was with Alex Rosenberg, who has provided very different perspectives and certainly sculpted the way I approach molecular biology today. It has been an absolute blast to bounce ideas off each other and learn from his experiences.

There are a handful of people who showed me the ropes early on, who are the reason I even decided to pursue a PhD. My very first research mentors, Harish Veeramani, with professor Michael Hochella introduced me to the research side of biology at a time where my pipetting skills were minimal at best. These two along with a postdoc who joined later, Jie Xu, laid down the foundation for my experimental design process. On research internships, I worked closely with Matthew Caporizzo while in Russell Composto's lab and with Steve Kennedy in David Mooney's lab. Both experiences exposed me to very different, yet equally exciting, fields of research.

I would like to thank all my collaborators outside the lab - David Peeler, Drew Sellers, Suzie Pun, Bosiljka Tasic, Zizhen Yao, Lucas Graybuck, Thuc Nguyen, Zach Thomson, Kimberly Aldinger, Kathleen Millen, Kaytlyn Gerbin, Tanya Grancharova, Ru Gunawardane, Kelly Paulson,

and Aude Chapuis. You are all a perfect example of the fantastic collaborative and translational environment that exists here in Seattle.

During my time in the PhD program, I was privileged to participate in two different training grants. First, I was on the NIH TL1 Translational Research that was sponsored through the Institute for Translational Health Sciences (ITHS). I thank JoAnne Whitney, Linda LeResche, and Patrick O'Keefe for organizing and running this program. I also participated in the NIH F32 Interdisciplinary Training in Research (IDTG) grant sponsored by the Fred Hutch Cancer Research Center. I thank Barry Stoddard and Anissa Barker for organizing and running this program.

I also want to thank Anna Kuchina, Paul Sample, Sumit Mukherjee, Matthew Hirano, Nick Bogard, Yue Zhang, Sifang Chen, Alex Rosenberg, and Richard Muscat for being awesome collaborators within the Seelig Lab, in addition to everyone else in Seelig Lab – Alberto Carignano, Sergii Pochekailov, Max Darnell, Alex Baryshev, Johannes Linder, Sunny Rao, Ban Wang, Erin Wilson, Sonya Volgeler, Arjun Khakar, Randolph Lopez, Ben Groves, and Gourab Chatterjee - who have kept science fun all these years.

I would like to thank my supervisory committee, Georg Seelig, James Carothers, Suzie Pun, and Barry Lutz, for providing great feedback during my qualifying, general, and final exams while also providing great general guidance throughout my time in the program.

Of course I cannot forget all of the phenomenal friends I've made here in Seattle. I will cherish all the adventures we've taken - from summiting mountains to hanging out at breweries to kayaking on the lake and doing everything else in between. Finally, I would like to thank my mom and dad for being continuously supportive of the life decisions I've made so far, and my fearless sister for constantly motivating me to do more in all aspects of life.

Table of Contents

.

Introduction

When I entered grad school, I wasn't sure what I wanted to work on, but I came with an overarching goal of trying to make an impact. I wanted whatever I worked on to be something that others – whether clinicians or researchers – would find legitimate use to further progress their work. I found myself focusing in on the genomics community, fortunate enough to work in a lab surrounded by highly skilled biologists, physicists, and computer scientists. Along the way I tried to learn the technical details in molecular and computational biology as well as broader scope of what the community needs were. I chose to focus on molecular biology methods and tools, where I saw many parallels to the development of computers: the tools that makeup a computer (i.e CPU, RAM, etc.) needed significant innovation before topics like machine learning could be fully realized - daily tasks that our laptops perform today were monumental findings just a few decades ago. Similarly, innovation in tools used during biological experiments can largely dictate the level of findings that can be achieved. The creation of molecular biology tools that enable higher throughput data generation or reveal new types of information will lead to exponential progress in our understanding of biology. It is my hope that this work will accelerate discovery for researchers and mark another step forward in the molecular biology technological landscape.

The advent of next generation sequencing has revolutionized molecular biology with the ability to read billions of molecules in parallel. As sequencing continuously increases in scale and drops in price, the genomics community is observing a shift in importance to library preparation. The difficulty now is not how to measure molecules, but how to extract important information from biological tissues and transform this into molecules that can be sequenced. In this work, I will discuss a number of technological developments that aim to accelerate the community's understanding of how to classify tissues and understand their functions.

A central goal of this work is to use sequencing data to understand biological phenotype. A majority of sequencing performed today is looking at the genome level – whether it is whole genome sequencing, exome sequencing, or a panel for specific genomic mutations. While valuable insights can certainly be gained at the DNA-level, this information can often fail to confidently describe biological function. For instance, say an unknown mutation is observed at the DNA level. This mutation could be the driver for a severe disorder or cause no biological effect whatsoever, but with no additional information it would be difficult to associate any biological (in)significance to this mutation.

The ideal measurement to infer biological phenotype would be at the protein level, as proteins are, in most cases, what actually carry out function. It is easy to measure some individual proteins with antibodies, but it is currently too difficult to measure hundreds, thousands, or tens of thousands of different proteins simultaneously. For most proteins, highly specific antibodies have yet to even be developed (*1*, *2*). For these reasons, the majority of my work has focused on transcriptomic measurements – the RNA level. Messenger RNA, or mRNA is essentially a proxy for protein expression. It is possible to capture mRNA representing all possible ~20,000 genes at the same time, making it attractive to measure all biological functions in unbiased fashion.

Briefly, the topics this work will cover are:

Scalable Single Cell RNA-seq. The first chapter discusses the development of a scalable single cell RNA-seq (scRNA-seq) assay we call SPLiT-seq. SPLiT-seq enables dramatic scaling in the numbers of cells that can be sequenced when compared to other scRNA-seq methods with no requirement of custom instrumentation. In the following chapter, we demonstrate the power of SPLiT-seq by profiling the entire brain and spinal cord of two mice. While previous studies were focused on specific regions of the brain, the scale at which we were able to operate allowed us to

examine developmental lineages and characterize cells types across multiple regions in the same experiment.

Sequence Specific Targeted Enrichment. While performing scRNA-seq provided a scalable *unbiased* approach to measure gene expression, there are cases when a *biased* approach is actually preferred. This third chapter will discuss the development of CleavR, a targeted enrichment strategy that allows for the selection of specific molecules from a pool of many others to make sequencing more efficient and raise the limit of detection for rare molecules.

Massively Parallel Report Assay for Splicing. A number of DNA mutations lead to aberrant splicing at the pre-mRNA level, leading to severe genetic disorders. The fourth chapter discusses the design and validation of a massively parallel reporter assay (MPRA) capable of screening hundreds of thousands of mutations for exon skipping incidence – a common source of aberrant RNA splicing.

Translation. I will conclude with a section discussing efforts made to get SPLiT-seq up and running into other labs. As my original mission when entering graduate school was to create an impact and enable the genomics community, this was particularly rewarding.

Chapter 1: Developing a Scalable Single Cell RNA Sequencing Technology

1.1: Motivation

Over three hundred years have passed since Leeuwenhoek first described living cells, yet we still do not have a complete catalogue of cell types or their functions. To do this, we need an efficient way to probe for biological functions in a sample and the ability to get single cell resolution. Addressing this first task would ideally involve measuring all proteins in a sample, but the need for a different antibody for every protein prohibits the ability to probe for the upwards of 20,000 different possible expressing proteins. Measuring all mRNA molecules, which are considered a proxy for protein expression and subsequent biological function, can be performed by using a single primer. Thus, RNA sequencing (RNA-seq) provides an unbiased measure of all possible biological functions in a sample.

Traditional RNA sequencing performs bulk measurements to derive a single RNA expression profile from tissues or large numbers of cells. Due to heterogeneity in biological samples, much of this information is convoluted because this RNA expression profile is the average of multiple profiles from different cell types. Recently, transcriptomic profiling of individual cells has emerged as an essential tool for characterizing cellular diversity (*3–5*). Single cell RNA-seq (scRNA-seq) methods have profiled tens of thousands of individual cells (*6–9*), revealing new insights about cell types within both healthy (*10–17*) and diseased tissues (*18–21*). Due to these advances, large initiatives such as the Human Cell Atlas launched in part by the NIH and the Chan Zuckerberg Biohub aim to use scRNA-seq to catalogue all cell types across the projected 37 trillion cells in the human body.

1.2: Overview of existing technologies

Most prominent scRNA-seq methods include drop-seq (*6*), InDrop (*7*), Seq-Well (*8*), and commercial platforms such as 10x Genomics (*9*). These methods all follow a similar concept, which relies on physically isolating individual cells in microwells or droplets before barcoding cDNA with a cell-specific tag. This intricate process of cell compartmentalization requires the use of complex microfluidic instruments which are costly and difficult to reproduce. Furthermore, for every new cell that is to be sequenced, a new compartment must be made with a new set of reagents, making the cost to scale in the number of cells linear.

1.3: Approach to a scalable alternative

To combat limitations of existing methods, we developed Split Pool Ligation-based Transcriptome sequencing (SPLiT-seq), a low-cost scRNA-seq method that enables transcriptional profiling of hundreds of thousands of fixed cells or nuclei in a single experiment (*22*). SPLiT-seq does not require partitioning single cells into individual compartments (droplets, microwells or wells), but relies on the cells themselves as compartments. The entire workflow before sequencing consists just of pipetting steps and no complex instruments are needed.

In SPLiT-seq, individual transcriptomes are uniquely labeled by passing a suspension of formaldehyde fixed cells or nuclei through four rounds of combinatorial barcoding. In the first round of barcoding, cells are distributed into a 96-well plate and cDNA is generated with an in-cell reverse transcription (RT) reaction using well-specific barcoded primers. Each well can contain a different biological sample – thereby enabling multiplexing of up to 96 samples in a single experiment. After this step, cells from all wells are pooled and redistributed into a new 96-well plate, where an in-cell ligation reaction appends a second well-specific barcode to the cDNA.

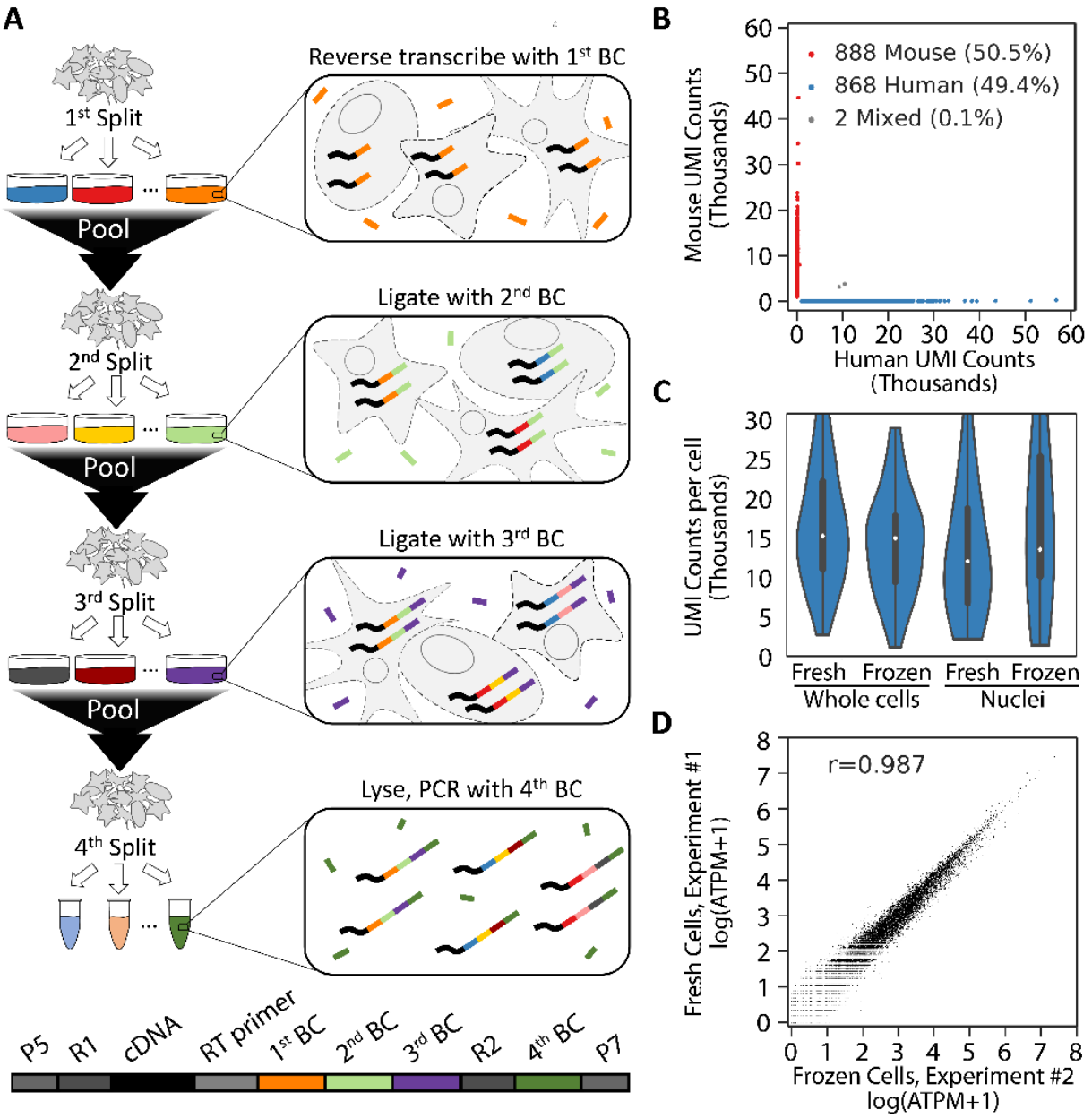


Fig. 1.1: Overview of SPLiT-seq. (A) Labeling transcriptomes with split-pool barcoding. In each split-pool round, fixed cells or nuclei are randomly distributed into wells and transcripts are labeled with well-specific barcodes. Barcoded RT primers are used in the first round. Second and third round barcodes are appended to cDNA through ligation. A fourth barcode is added to cDNA molecules by PCR during sequencing library preparation. The bottom scheme shows the final barcoded cDNA molecule. (B) Species mixing experiment with a library prepared from 1,758 whole cells. Human UBCs are blue, mouse UBCs are red, and mixed-species UBCs are gray. The estimated barcode collision rate is 0.2%, whereas species purity is >99%. (C) UMI counts from mixing experiments performed with fresh and frozen (stored at -80°C for 2 weeks) cells and nuclei. Median human UMI counts for fresh cells: 15,365; frozen cells: 15,078; nuclei: 12,113; frozen nuclei: 13,636. (D) Measured gene expression by SPLiT-seq is highly correlated between frozen cells and cells processed immediately (Pearson-r: 0.987). Frozen and fresh cells were processed in two different SPLiT-seq experiments.

The third-round barcode, which also contains a unique molecular identifier (UMI), is then appended with another round of pooling, splitting, and ligation. After three rounds of barcoding, the cells are pooled, split into sublibraries, and sequencing barcodes are introduced by PCR. This final step provides a fourth barcode, while also making it possible to sequence different numbers of cells in each sublibrary. After sequencing, each transcriptome is assembled by combining reads containing the same four-barcode combination (Fig. 1.1A, Fig. S1A).

Four rounds of combinatorial barcoding can yield 21,233,664 barcode combinations (three rounds of barcoding in 96-well plates followed by a fourth round with 24 PCR reactions) - enough to uniquely label over 1 million cells. Even larger numbers of barcode combinations can be achieved by performing experiments in 384-well plates or through additional rounds of barcoding (Fig. S1B). In addition, by performing the first step in a 384-well plate, up to 384 different biological samples could be combined in a single experiment.

1.4: Validating a newly developed method

To test SPLiT-seq's ability to generate uniquely barcoded cells (UBCs), a species-mixing experiment was performed. It is possible to computationally differentiate between RNA molecules deriving cells of two different species based on RNA alignment. If the RNA from cells that SPLiT-seq outputs aligns to both species, it would indicate that two or more cells are being perceived as a single cell from the experiment. However, if RNA from all cells that SPLiT-seq outputs aligns only to one of the two species, it would indicate that true single cell RNA sequencing is being achieved.

Cells from one mouse and two human cell lines (NIH/3T3, HEK293, and Hela-S3) were mixed, formaldehyde fixed, and used SPLiT-seq to generate a scRNA-seq library with 1,758

UBCs. The library was sequenced and reads were aligned to a combined mouse-human genome. 99.9% of the UBCs were unambiguously assigned to a single species (>90% of reads aligned to a single genome) with the remaining 0.1% of UBCs representing barcode collisions between mouse and human cells (Fig. 1.1B). At saturating read coverage (>500,000 reads per cell), a median of 15,365 UMIs and 5,498 genes per human cell and 12,243 UMIs and 4,497 genes per mouse cell was identified. The species purity in both human and mouse UBCs was high: 99.6% of reads in human UBCs and 99.0% of reads in mouse UBCs aligned to their respective genomes. SPLiT-seq experiments were repeated with freshly prepared nuclei as well as nuclei and cells that had been preserved at -80°C for two weeks. In all samples, similar numbers of transcripts and genes per cell were detected (Fig. 1.1C, Fig. S2, Table S1). Gene expression was highly correlated between preserved and freshly prepared cells (Fig. 1.1D, Fig. S2, Pearson-r: 0.987) as well as between cells and nuclei (Fig. S2, Pearson-r: 0.952). Examining gene and UMI detection at different sequencing depths revealed that the sensitivity of SPLiT-seq is comparable to droplet-based scRNA-seq methods (Fig. S3).

1.5: Conclusion

SPLiT-seq's compatibility with fixed cells and fixed nuclei overcomes challenges faced by other scRNA-seq methods. Fixation can reduce perturbations to endogenous gene expression during cell handling (*23*) and makes it possible to store cells for future experiments. Moreover, the use of nuclei bypasses the need to obtain intact single cells, which can be challenging for many complex tissues. SPLiT-seq's compatibility with formaldehyde-fixed nuclei suggests it may be used to profile single nuclei from formalin-fixed, paraffin-embedded tissue (*24*).

SPLiT-seq enables flexible and scalable cell and sample multiplexing. The use of the first-round barcode as a sample identifier makes it possible to profile a large number and variety of

samples in parallel, thus minimizing batch effects. As the number of unique barcodes grows exponentially with the number of barcoding rounds, larger numbers of cells than presented here could be processed by adding a fifth barcoding round or by switching to a 384-well plate format. Although for such large cell numbers sequencing cost may currently be forbidding, it is easy to imagine extended applications, such as targeted sequencing of gene panels, which would even now benefit from very large cell numbers and only require shallow sequencing depth.

Our hope is that the increased scale and accessibility provided by the low cost and minimal equipment requirements of SPLiT-seq will further accelerate the widespread adoption of scRNA-seq. To facilitate adoption, a website with the detailed protocol (Supplemental materials) and answers to frequently asked questions is available for other researchers. We also transformed SPLiT-seq into a kit form to further enable other labs on getting the method up and running. Please refer to chapter 5 to learn more about efforts in translating SPLiT-seq into other labs.

Chapter 2: Profiling Development in the Central Nervous System

2.1: Overview of experimental process

SPLiT-seq was used to profile nuclei from the developing brain and spinal cord of postnatal day 2 and 11 (P2 and P11) mice. The first round of barcoding assigned identifiers for the P2 brain, P2 spinal cord, P11 brain, and P11 spinal cord samples (Fig. 2.1A, Fig. S4). In total, four rounds of barcoding (48 x 96 x 96 x 14) generated over 6 million distinct barcode combinations, making it possible to process hundreds of thousands of nuclei in a single experiment with minimal barcode collisions (2.5% expected collisions for 150,000 nuclei).

To test how many transcripts SPLiT-seq detects within nuclei from the central nervous system, deep sequencing on a sublibrary containing only 131 nuclei was performed, detecting 4,943 UMIs and 2,055 genes per nucleus (UMI duplication: 95%). The rest of the library was sequenced at lower depth, resulting in a median of 677 genes and 1,022 UMIs per nucleus (UMI duplication: 58%) (Table S2). Low-quality transcriptomes were removed from analysis (*25*), yielding 156,049 single-nucleus transcriptomes (74,862 P2 brain; 7,028 P2 spinal cord; 58,573 P11 brain; 15,586 P11 spinal cord).

Unsupervised clustering grouped transcriptomes into 73 distinct clusters (*25*) (Tables S3-S5), which were visualized by t-Distributed Stochastic Neighbor Embedding (t-SNE, Fig. 2.1A). Each of these 73 clusters was assigned to a cell class on the basis of expression of established marker genes (Fig. 2.1B). Neurons accounted for 79% of the profiled transcriptomes (54 clusters), with most clusters expressing *Meg3*.

2.2: Non-neuronal cell types

The 27,096 non-neuronal transcriptomes spanned 19 different clusters, each assigned to a specific cell type. Four astrocyte types (Fig. 2.1C) accounted for 50% of all non-neuronal nuclei

(n=13,481). Oligodendrocytes (6 types, n=4,294) and oligodendrocyte precursor cells (OPC, 1 type, n=5,793) formed the second most abundant population. Further analysis identified two vascular and leptomeningeal cell (VLMC) types (Fig. S5A), endothelial cells, smooth muscle cells

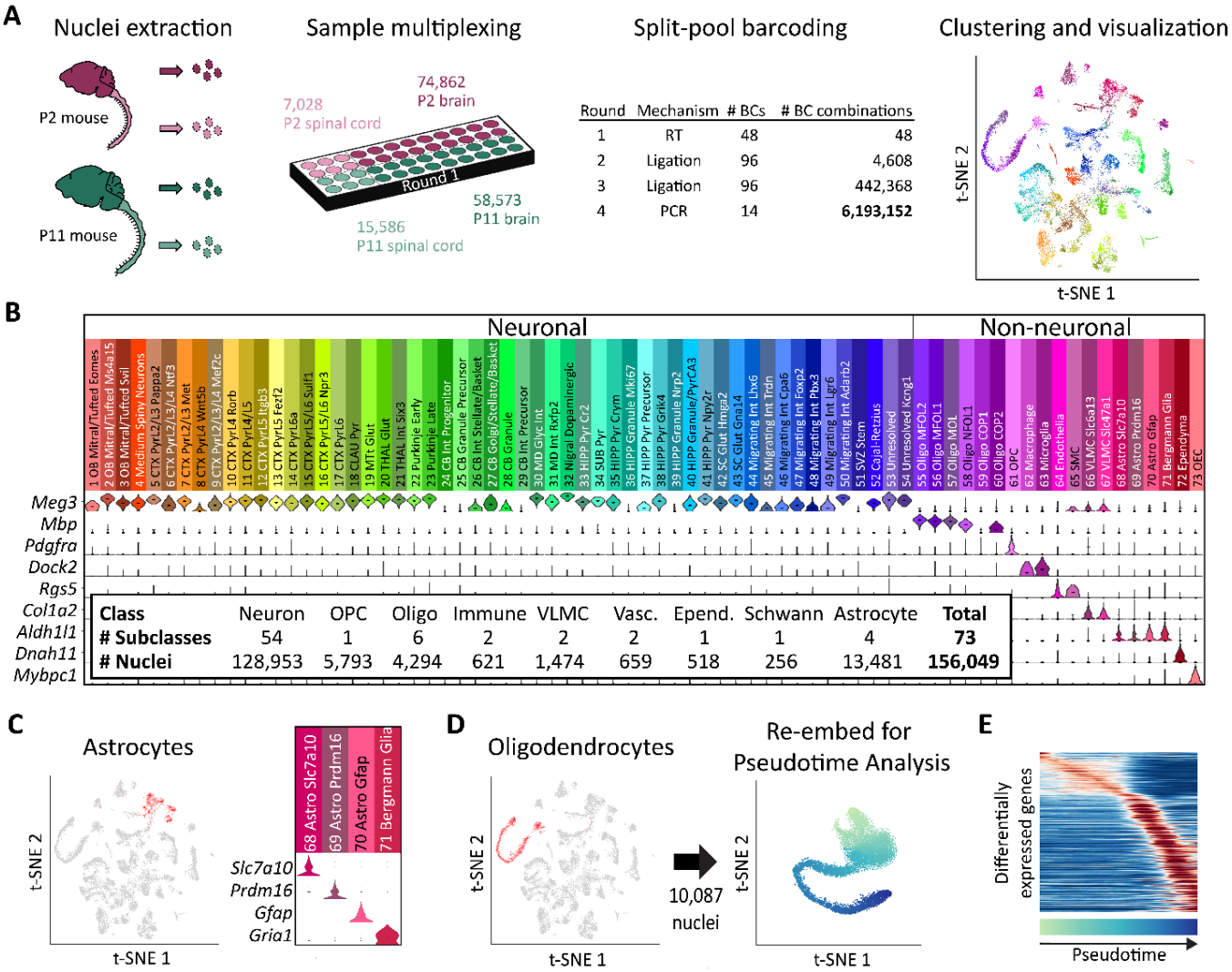


Fig. 2.1: Single-cell transcriptome landscape of postnatal brain and spinal cord development by SPLiT-seq. (A) Over 150,000 nuclei from P2 and P11 mouse brains and spinal cords were profiled in a single experiment employing over six million barcode combinations. Transcriptomes were clustered and then visualized using t-SNE. Cells are colored according to cell type. Each cluster was downsampled to 1,000 cells for visualization. (B) A total of 73 distinct clusters were assigned to nine cell classes based on expression of established markers. The violin plots show marker gene expression in each cluster. (C) Astrocyte clusters are highlighted in red in the t-SNE. The violin plots show markers that are differentially expressed between astrocyte subtypes. (D) Seven OPC and oligodendrocyte clusters (containing 10,087 nuclei) colocalized in the original t-SNE (highlighted in red), forming a lineage. Cells from these clusters were re-embedded with t-SNE. (E) The heatmap shows genes expressed differentially across pseudotime in the oligodendrocyte lineage.

(Fig. S5B), microglia, macrophages (Fig. S5C) (*26*, *27*), ependymal cells, and olfactory ensheathing cells.

Previous work has observed that t-SNE can order cells in 2D space according to stages of differentiation (*12*). Moving through t-SNE space along the path of differentiation can then be viewed as moving through "pseudotime"(*28*). As oligogenesis spans the first two postnatal weeks of murine development (*29*), we asked whether the oligodendrocyte and OPC clusters might reflect a continuous developmental trajectory. When examining the oligodendrocyte clusters, we found that they formed an overlapping elongated shape in the t-SNE visualization. OPCs and oligodendrocytes from the P2 mouse, were enriched at one end of the structure while oligodendrocytes from the P11 mouse were enriched at the opposite end (Fig. S6), indicative of a lineage (*25*, *28*).

A more thorough analysis was performed on this putative lineage. To ensure that our ordering of oligodendrocytes was determined exclusively by their relationship to other oligodendrocytes, rather than all cells, we re-embedded only transcriptomes within these seven clusters with t-SNE (Fig. 2.1D, Fig. S7A). We calculated the moving average of gene expression in the resulting pseudotime ordering (Fig. 2.1E, Fig. S8). Analysis of these expression patterns confirmed that proliferating OPCs segregated to one end of the t-SNE, whereas mature oligodendrocytes segregated to the opposite end (Fig. S7B). We also detected previously reported intermediate stages of oligodendrocyte development, with the order of gene expression across pseudotime nearly identical to the one defined previously (*12*) (Fig. S7C, Spearman-r: 0.94). When analyzing spinal cord and brain derived cells separately, we found more mature oligodendrocytes in the spinal cord than in the brain (Fig. S7D), indicating that oligodendrocyte maturation occurs earlier in the spinal cord.

2.3: Neuronal cell types

Using known gene markers, we were able to assign most neuronal clusters to specific cell types (*25*). While some clusters corresponded to abundant cell types, such as cerebellar granule cells (CGCs), others mapped to rare and often less characterized cell types, such as mitral/tufted cells. Previously characterized regional markers were used to assign the majority of clusters to a specific region of the brain (*30*) (Fig. 2.2A). Regional assignments were validated with RNA in situ hybridization (ISH) from the Allen Institute's Developing Mouse Brain Atlas (Allen DMBA) (*31*). Specifically, we generated composite ISH maps by averaging across the five most highly enriched genes from each of our clusters (Tables S6, S7). For clusters primarily containing P2 or P11 nuclei, we used the P4 or P14 atlases, respectively. The resulting composite maps confirmed the high regional specificity of most types (Figs. 2.2B, S9 and S10). Cortical pyramidal neuronal types could be further assigned to specific layers using marker genes (Fig. 2.2C) (*10*, *11*).

In the hippocampus, immature granule cells in the dentate gyrus give rise not only to mature granule cells, but also to pyramidal neurons (*32*). This process is one of two instances of neurogenesis that continues into adulthood (*33*), but little is known about the underlying transcriptional program. We determined that three neuronal cell types from the hippocampus likely constituted a developmental trajectory (*25*). Analysis of only these transcriptomes with t-SNE revealed a clear branching structure (Fig. 2.2D, Fig. S11A). The transcription factor *Prox1*, suspected to be necessary for granule cell identity (*34*), was exclusively expressed in one branch, whereas genes known to be specific to CA3 pyramidal neurons such as *Spock1* (*35*) were expressed exclusively in the other branch. Markers of dividing neuronal progenitors were expressed before the branching point and genes in the Slit-Robo signaling pathway were differentially expressed between the two lineages (Fig. S11B). We used these data to identify specific temporal dynamics

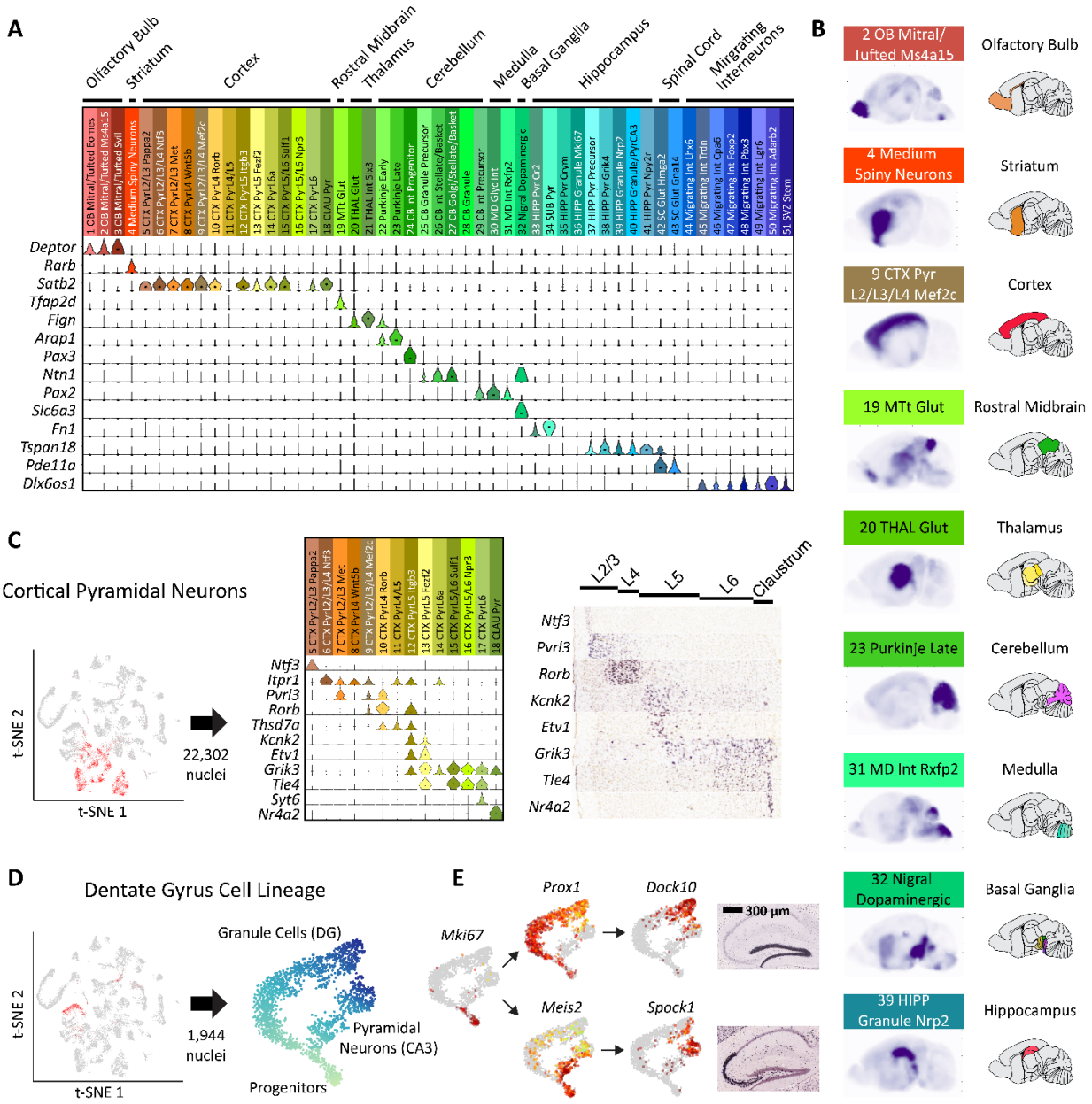


Fig. 2.2: Neuronal clusters exhibit regional specificity. (A) Marker gene expression was used to map neuronal clusters to specific brain regions. (B) Sagittal composite RNA ISH maps for nine representative clusters from distinct areas. For each cell type, ISH intensities were averaged from the Allen DMBA across the top five differentially expressed genes. (C) Types of pyramidal neurons in the cortex display layer-specific enrichments according to marker genes: cortical pyramidal neurons are highlighted in red in the t-SNE. Expression of example marker genes in pyramidal clusters is shown in the middle and corresponding available RNA ISH results on the right. (D) Three clusters constitute a developmental trajectory in the hippocampus. Re-embedding these clusters highlights the branching of the two differentiation trajectories in pseudotime. (E) Expression of differentiation marker genes is overlaid on the t-SNE. RNA ISH maps (Allen DMBA) show the regional specificity of granule cell and pyramidal neuron markers.

pyramidal cell differentiation (Fig. 2.2E, Fig. S12).

2.4: A closer look into the cerebellum

The cerebellum accounts for only 9% of the brain mass in adult mice, but contains nearly 85% of all neurons (*36*). Despite the wide range of functions performed by the cerebellum, many of the gene expression programs driving development of cerebellar cell types remain unknown. We identified the four main cerebellar neuronal types (Fig. 2.3A): Purkinje cells, Golgi cells, stellate/basket cells, and CGCs. Two types of Purkinje cells (Fig. 2.3B) were segregated primarily by age (P2 vs P11), and did not form a continuous trajectory in t-SNE but rather two clearly segregated clusters. The absence of cells at intermediate stages of maturation suggests that Purkinje cell development may be more synchronous than other processes of neurogenesis captured by our dataset.

CGCs, the most numerous type of neuron in the brain (*37*), drive the postnatal foliation of the cerebellar cortex by migrating from the external granule layer (EGL), through the molecular layer (ML) and the Purkinje cell layer (PcL) to the internal granule layer (IGL) (*38*, *39*). We created a pseudotime ordering of 15,360 CGCs (Fig. 2.3C, Fig. S13) and measured gene expression across this lineage. We defined genes with specific expression at different points in pseudotime (Fig. S14), and then used RNA ISH to map these genes to layers of the developing cerebellar cortex. Genes ordered from early-to-late in pseudotime were progressively expressed from outer-to-inner layers, consistent with the known direction of CGC migration (Fig. 2.3D). Our analysis revealed previously unknown pseudotime and layer specific gene expression patterns within pathways related to axonal development and neuronal migration (Fig. S15).

The question of whether all cerebellar inhibitory interneurons arise from the same progenitor population has been a point of contention (*40*). Early hypotheses proposed that

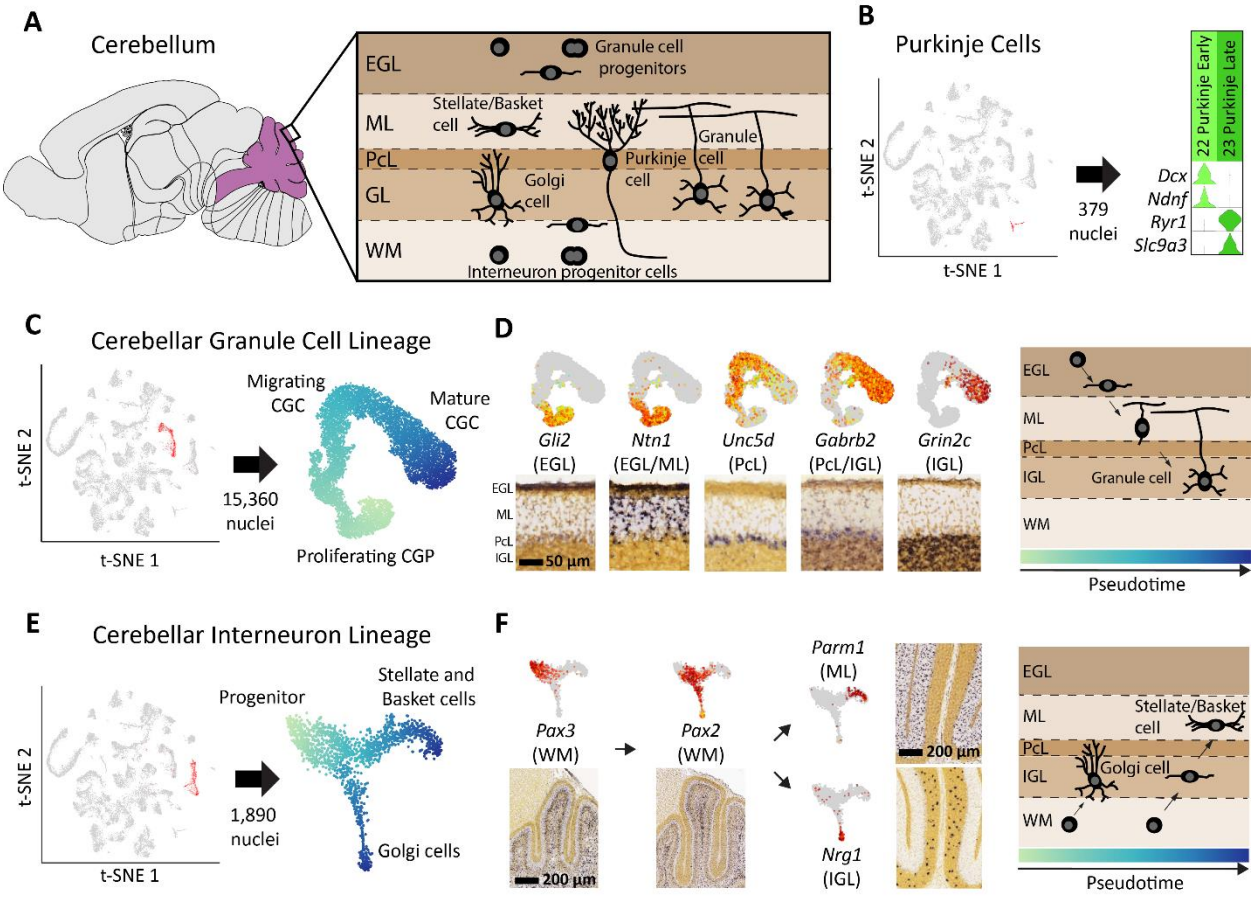


Fig. 2.3: Neuronal differentiation trajectories in the cerebellum revealed by SPLiT-seq. (A) Major cell types and their locations in the cerebellum. (B) Two types of Purkinje cells with distinct gene expression programs were identified. Early Purkinje cells are primarily found in the P2 brain and late Purkinje cells in the P11 brain. (C) t-SNE re-embedding of 15,360 nuclei suggests a pseudotime ordering from proliferating, to migrating, to mature CGCs. (D) Expression of marker genes is overlaid on the t-SNE, and the corresponding RNA ISH from Allen DMBA is shown below. Marker genes associated with different layers of the cerebellum are expressed at different points in pseudotime. Gene expression order is consistent with ordering of the physical layers. RNA ISH maps confirm regional specificity of marker genes. (E) t-SNE re-embedding of 1,890 nuclei reveals a branching differentiation trajectory. Progenitors can either become Golgi cells or stellate/basket cells. (F) Markers for progenitors and mature cell types are expressed at different points in pseudotime and have layer specificity.

stellate/basket cells originated from precursors in the EGL, whereas Golgi cell precursors resided in the ventricular epithelium (*41*). Later evidence indicated that these two interneurons shared a

common precursor in the cerebellar white matter (*42*, *43*). However, the molecular profile of the inhibitory neuron lineage in the cerebellum remains largely unknown.

We found a cerebellar inhibitory interneuron lineage (1,517 cells, Fig. 2.3E, Fig. S16A) with a shared progenitor branching into either Golgi or stellate/basket cells (Fig. S17). This lineage includes a known precursor cell type expressing *Pax2* (*42*), but also a previously unknown, earlier precursor expressing *Pax3* (Fig. 2.3F). RNA ISH analysis suggests that this Pax3+ precursor is located deep within the cerebellar white matter. Moreover, we found that stellate/basket cells expressed genes specific to the molecular layer, whereas Golgi cells expressed genes specific to the granule cell layer (Fig. 2.3F, Fig. S18). The distribution of P2 and P11 nuclei within the lineage clearly demonstrated that the maturation of Golgi cells was well underway by P2 and complete by P11 (Fig. S16B). In contrast, stellate/basket cells had not begun to differentiate at P2 and were still not fully mature by P11. These results indicate that the same molecularly defined precursor gives rise to two distinct interneurons at different stages of development.

2.5: Examining the spinal cord

The original clustering was dominated by cells in the brain, and many spinal cord cells did not segregate into well-defined clusters (Fig. S19). To resolve more cell types in the spinal cord, we selected all the nuclei originating from the spinal cord and re-clustered them (*25*), resulting in 44 clusters: 14 non-neuronal types (12 of which were also found in the brain) and 30 neuronal types (Fig. 2.4A and Tables S8-S10). We identified 11 different types of GABAergic neurons, of which several were also glycinergic (Fig. 2.4B). One GABAergic type was identified as cerebrospinal fluid-contacting neurons (CSF-cNs) (*44*), with the other ten types corresponding to inhibitory interneurons. Glutamatergic interneurons accounted for 15 additional types. We also identified two clusters of cholinergic motor neuron types (alpha and gamma) (*45*). To date, known

markers exist only for gamma motor neurons (*Esrrg* and *Esrrb*) (*46*), however, we identified specific markers for both alpha and gamma neurons (Fig. 2.4C).

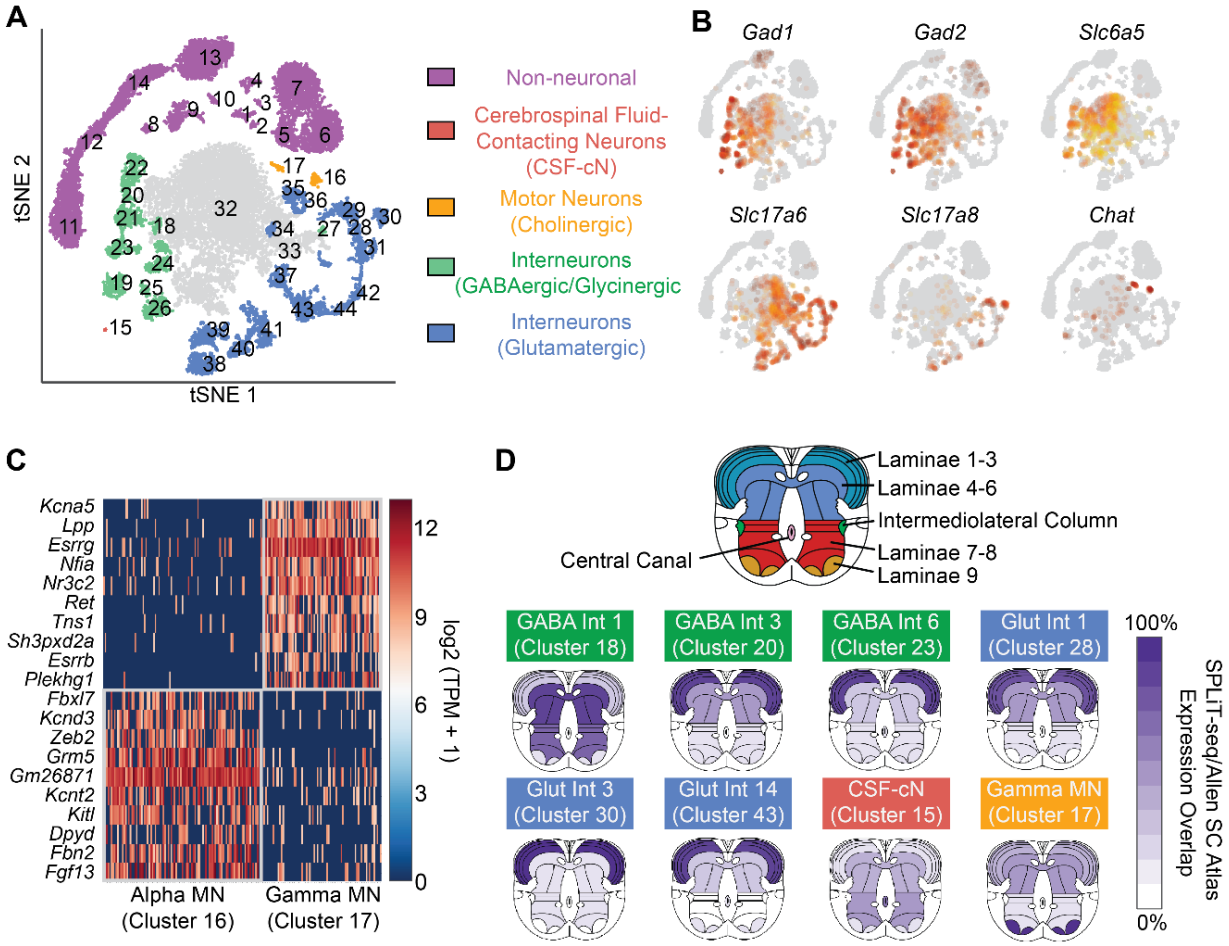


Fig. 2.4: Gene expression patterns and spatial origin of cell types in the spinal cord. (A) Re-clustering spinal cord nuclei resulted in 30 neuronal and 14 non-neuronal clusters. (B) GABAergic neurons were defined by expression of *Gad1* and *Gad2*. A subset of GABAergic neurons are also glycinergic, based on expression of *Slc6a5*. Glutamatergic neurons were defined by expression of VGLUT1 (*Slc17a6*), whereas cholinergic motor neurons express *Chat*. (C) Novel gene markers distinguish gamma motor neurons from alpha motor neurons. (D) Inferred spatial origin of neuronal clusters within the spinal cord. We analyzed the Allen Spinal Cord Atlas expression patterns of the top ten enriched genes in each cluster. Dark purple indicates expression of all ten genes in the given region, while white indicates none of the ten genes were expressed in the given region.

To infer the spatial origin of neuronal types in the spinal cord, we identified the ten most enriched genes in each type according to our snRNA-seq data and created composite ISH maps based on the Allen Mouse Spinal Cord Atlas (*47*) (Fig. 2.4D, Fig. S20). Some interneuron subtypes

appeared to originate primarily from laminae 1-3, with others originating from laminae 4-6. We found both inhibitory and excitatory neurons in each region. Motor neurons expressed genes found in laminae 9, while CSF-cNs were the only neuronal type expressing genes found in the central canal. These data allowed us to create an atlas of gene expression in the early spinal cord, providing a rich resource for further understanding development of the central nervous system.

2.6: Conclusion

In this work, hundreds of thousands of cells were profiled using only basic laboratory equipment with a library preparation cost of ~$0.01 per cell (Fig. S21, Table S11). In the analysis of more than 150,000 single-nucleus transcriptomes from two early postnatal stages, 69 types of cells in the brain and 44 types in the spinal cord were identified. Many new molecular markers for specific cell types and a comprehensive gene expression profile for four different developmental lineages were defined, providing an excellent resource for neuroscientists moving forward.

Chapter 3: Developing a Targeted Sequencing Technology

3.1: Motivation

The demand for high throughput single cell sequencing is expected to expand in the coming years as human cell atlas projects grow and researchers begin to bring scRNA-seq into the clinic. We will need technologies capable of scaling up single cell analysis both in cells and samples at a low cost. While SPLiT-seq has aided in this front on the library preparation side, there is another large cost factor that remains unsolved: sequencing itself. For every new cell that is prepared for sequencing, there are a new set of molecules that must be sequenced. Therefore, the cost to sequence the molecules from a single cell RNA-seq library still scales linearly as more cells are added.

One of the key features of RNA-seq is that it is capable of measuring expression across all genes, providing a comprehensive and unbiased view. However, in many cases we do not need an unbiased view. Some scientists already know what genes are important to solve their research questions or to diagnose a patient with a certain condition. Increasing work is being performed to find key genes such as transcription factors and drivers of gene modules whose expression can guide the inference of the expression of many other genes (*48*, *49*). In some cases, researchers only want to capture a few specific molecules from each cell, but need those types of information from millions of cells. Examples of this include capturing T cell or B cell receptor (TCR/BCR) sequences from immune cells (*50*, *51*), or examining areas of the transcriptome or genome known to carry a high mutational burden or signature for disease (*52–54*). These cases can be further exemplified by flow cytometry, where researchers routinely profile millions of cells at a low cost. However, flow cytometry is only capable of capturing expression from roughly 12 markers. This technology is often used in scenarios where hundreds of markers are actually needed to fully

explain biological function or understand diagnosis of a disease. Furthermore, flow cytometry is not able to capture sequence-based information such as single nucleotide polymorphisms, TCR, or BCR information as the detection is based only on fluorescence rather than sequencing (*55*, *56*).

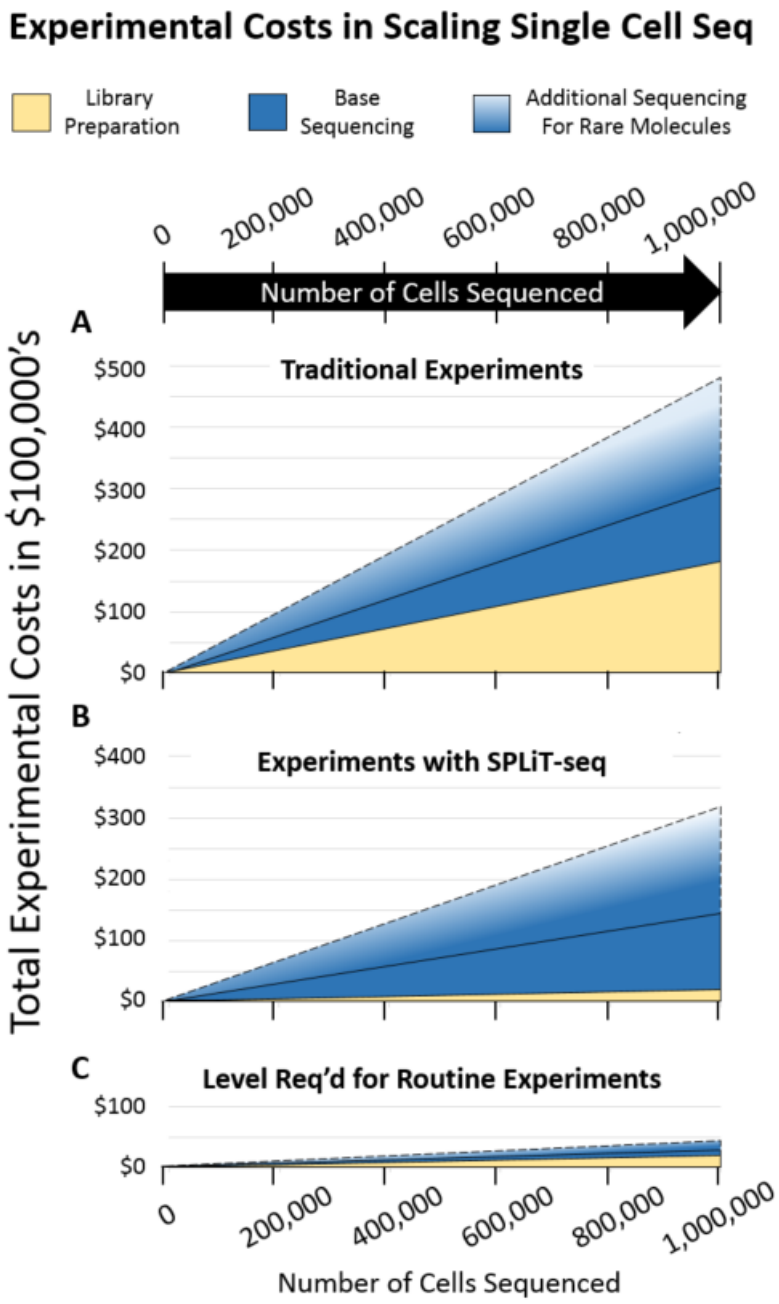


Figure 3.1: Total experimental cost estimates as a function of cells sequenced in single cell RNA-seq. Plots encompass the cost to generate a scRNA-seq library (yellow), cost to sequence sufficient molecules where data can be correctly interpreted and analyzed (solid blue), and cost for additional sequencing, which would be needed to extract higher quality data or information from rare molecules (blue gradient). (A) Traditional technologies: A 1M cell experiment ranges between $300,000 and $480,000 where both library preparation and sequencing are large contributors to cost. (B) Incorporating SPLiT-seq: Cost for 1M cell experiment drop to $140,000-$300,000, with majority deriving from sequencing. (C) What is needed for routine high throughput experiments: The sequencing costs must be greatly reduced to bring a 1M cell experiment to below $50,000.

Acquiring a single cell RNA-seq library of hundreds of thousands of cells on the current most common commercial scRNA-seq platform, followed by sequencing that library on the current

most scalable sequencing platform leads costs that are in the many hundreds of thousands of dollars (Fig. 3.1A, Supplementary Text). SPLiT-seq dramatically changes the library preparation cost when scaling these experiments to millions of cells but has no substantial effect on sequencing cost (Fig. 3.1B).

Given that many experiments do not require the sequencing of all genes captured by RNA-seq, the next question is are we able to selectively enrich for molecules of interest and effectively deplete insignificant molecules before a library is ever sequenced. This would reduce the number of molecules required to sequence per cell and therefore lead to cost savings (Fig. 3.1C). Additionally, rare transcripts that would typically go undetected could now be targeted and quantified. There are two traditional forms of technologies that can perform this type of target enrichment on molecules: multiplex PCR and bead-based probe hybridization. We will first explore these technologies before introducing a conceptually orthogonal method we have developed for target enrichment.

3.2: Overview of existing technologies

One enrichment method is multiplex PCR. This encompasses a PCR reaction where two primers (forward and reverse) are added to the reaction for each region of interest (ROI). Exponential amplification of ROIs can bring the representation of all other molecules to very low levels that are virtually undetectable when sequencing. Multiplex PCR can easily achieve thousands fold enrichment of ROIs when there are very few regions being targeted. However, there are significant design considerations when targeting more than three or four ROIs at a time. The specificity of PCR relies heavily on the melting temperature of the primers being used. While easy to design a few primers with the same melting temperature, it becomes increasingly difficult to design multiple primer sets that are compatible with one another (*57*). As a consequence, most

multiplex PCR methods set lower annealing temperatures to accommodate the variability in primer melting temperatures. These low annealing temperatures lead to mis-priming of unintentional targets that amplify exponentially in addition to chimeric products (*58, 59*). While some groups have designed multiplex PCR reactions with tens to hundreds of primers through heavy optimization, the reaction is not modular as the introduction of a new primer sets to these reactions often require additional optimization.

A second method commonly used is bead-based probe hybridization. Here, single stranded DNA oligonucleotides (ssDNA) that are complimentary to a target region are attached to paramagnetic beads, typically through a biotin streptavidin interaction (*60, 61*). Next, the sequencing library is melted and introduced to the ssDNA containing beads. The goal here is that the ssDNA probes on the beads will hybridize to the target regions of the single stranded sequencing library. Once that occurs, the beads can be pulled down with a magnet where the supernatant containing molecules that are not desired are removed. Most bead-based methods require an overnight incubation for hybridizations, making the protocol fairly long. These methods perform relatively well when enriching targets that consist a few percent of the entire library (i.e. exome sequencing), but are not as effective in capturing very rare transcripts that represent well below 1% of the population, largely due to non-specific binding to beads (*62*).

3.3: Experimental approach

To overcome limitations to current methods, we developed CleavR (cleavage by Rnase), a fast and modular targeted enrichment method that leverages the highly active and specific nature of RNases – enzymes capable of cutting or depleting RNA. Starting material for the method is an amplified sequencing library, whose universal PCR adapters have riboguanine (riboG) bases throughout the universal PCR adapter (Fig. 3.2A). This can be easily incorporated while doing the

original amplification of the sequencing library by using RNA-DNA hybrid primers. This library is then added to a mixture of a polymerase, T1 Rnase, and a series of capture primers. A single capture primer is needed for each type of molecule, such as a specific gene, that is to be enriched. This capture primer can be designed much like any primer set for PCR, with exception that only the forward primer is needed here.

This mixture is then subjected to a thermocycling protocol. First, all double stranded DNA are melted into single strands by heating the mixture to 95˚C (Fig. 3.2B). Next, the temperature is lowered to an annealing temperature compatible with the melt temperature of the capture primers. For capture primers in our study, we used an annealing temperature of 55˚C. The capture primer will anneal to the complimentary sequence on the strand where enrichment is desired (Fig. 3.2C). The temperature is then raised to 72˚C to promote polymerase extension of the annealed capture primer (Fig. 3.2D). Importantly, one of the universal adapters on this enrichment strand will become double stranded after the extension. The riboG base in this universal adapter has also become double stranded through this extension step. The thermocycling protocol deviates from a traditional PCR cycle by lowering the temperature to 37˚C where T1 Rnase is active. T1 Rnase is a highly active enzyme that cleaves single stranded riboG bases with high specificity. Because riboGs on the molecule that we want to enrich for is protected via the double stranded universal adapter, T1 Rnase can be deployed to attack all single stranded universal adapters that remain by cleaving their riboG base (Fig. 3.2E).

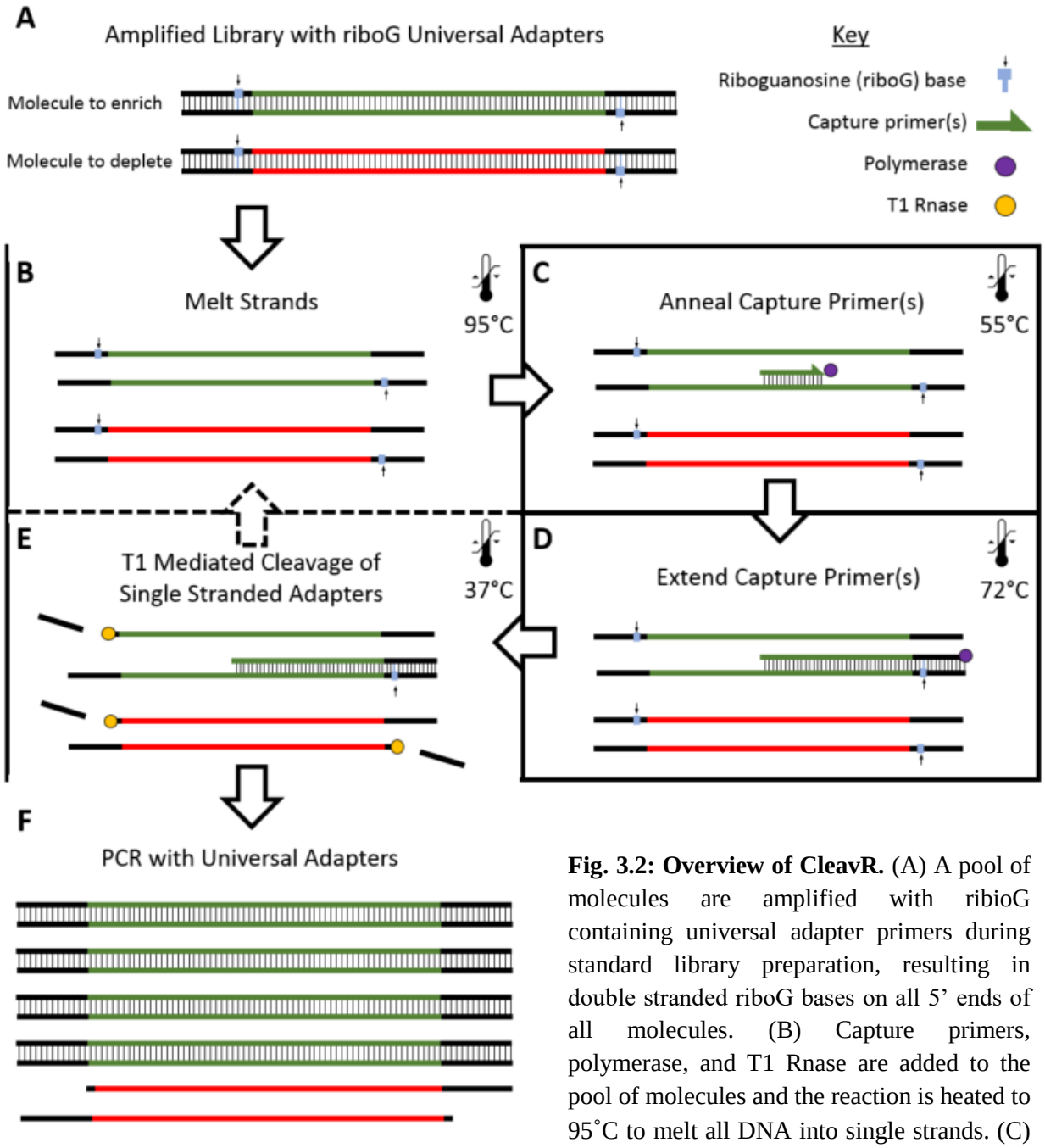


Fig. 3.2: Overview of CleavR. (A) A pool of molecules are amplified with ribioG containing universal adapter primers during standard library preparation, resulting in double stranded riboG bases on all 5' ends of all molecules. (B) Capture primers, polymerase, and T1 Rnase are added to the pool of molecules and the reaction is heated to 95˚C to melt all DNA into single strands. (C) The reaction temperature is lowered to an appropriate annealing temperature where capture primers bind to their complimentary target. (D) The temperature is increased allowing polymerase to extend the capture primer to end of the molecule, forming a double stranded riboG base at the end of the molecule where capture primers bound. (E) Reaction temperature is lowered to 37˚C, initiating T1 Rnase cleavage of remaining single stranded RiboG bases. After this step, the reaction can undergo another cycle where the strands are melted again to restart the process. (F) A PCR reaction using the universal adapter primers seletivley amplifies molecules that still contain a universal adapter region. Molecules whose universal adapter region was cleaved do not amplify and are effectively depleted from the library.

Due to potential errors in mis-priming of capture probes during a single step of extension, this thermocycling process can be performed more than once to ensure that molecules that we do not wish to enrich for have their universal adapter sequenced cleaved. Once this step is completed, this mixture undergoes a normal PCR reaction where primers complimentary to the universal adapters are used (Fig. 3.2F). Because the molecules we want to enrich for should be the only ones with intact universal adapter sequences, these are the only molecules that are amplified. After exponential amplification of PCR is completed, the selected molecules will resemble the majority of the pool of molecules.

It is easy to imagine failure modes in CleavR if the T1 Rnase and polymerase were always active. For instance, all universal adapters on strands, including those on enrichment strands, are single stranded during the melting step at 95˚C. If T1 Rnase were active at this point, all strands would be cleaved, rendering the method useless. However, T1 Rnase is known to denature at temperatures that are higher than approximately 50˚C. This protein also has very high conformational stability, enabling the enzyme to renature to a functional state when temperatures goes below this approximate 50˚C mark (*63*, *64*). This effectively creates a bi-functional switch for T1 Rnase that is controlled by temperature. While not as extreme, there could also be concern of the polymerase exhibiting extension of mis-primed capture probes at low reaction temperatures such as 37˚C. To combat this issue, we can use a reversible hot start polymerase. This polymerase comes loaded with an aptamer that will reversibly bind and inactivate the enzymatic activity at temperatures lower than approximately 45˚C (*65*). This essentially forms a similar temperature controlled bi-functional switch to that of the T1 Rnase with exception that the polymerase is only active at temperatures above 45-50˚C while the T1 Rnase is only active at temperatures below 45-50˚C.

3.4: Validating Sequenced-Based Enrichment of Targets

To evaluate CleavR's ability to enrich for specific molecules based on sequence, a control experiment was performed. Two DNA strands of different length were amplified using primers with riboG bases in the universal adapter region. The first strand derived from an ampicillin resistance gene section of a plasmid (amp strand) that was 441 bases long. The second strand derived from a hygromycin resistance gene section of a plasmid (hygro strand) that was 794 bases long. Capture primers were designed for amp and hygro strands that were respectively 102 and 84 bases upstream from the end of the universal adapter region. Initial experiments comprised of mixing the amp and hygro strands at a 1:1 ratio followed by two enrichment reactions: one containing amp capture primers and another containing hygro capture primers. When CleavR was performed, the reaction containing amp capture primers yielded a strong band at around 441 bases with no detectable band around 794 bases (Fig 3.3). Furthermore, the reaction containing hygro capture primers yielded a strong band at around 794 bases with no detectable band around 441 bases (Fig 3.3). These results indicate that the method was successful at enriching for one strand over the other based on the sequence of the capture primer.

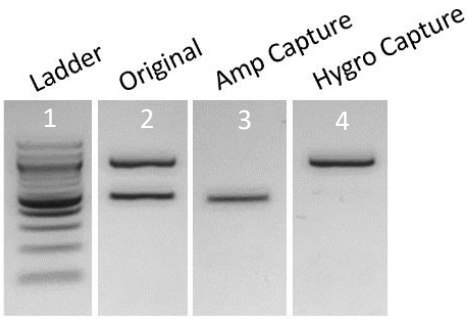


Fig. 3.3: Pilot experiment of CleavR. Short DNA strands (Amp) and long DNA strands (Hygro) are mixed at a 1:1 ratio (Lane 2). When attempting to capture Amp strands, only the lower band is visible (Lane 3) while only the larger band appears when attempting to capture Hygro strands (Lane 4).

Next, we sought out to determine the fold enrichment that was being achieved through the assay. The first experiment demonstrated some level of enrichment but was not able to quantitatively determine the fold enrichment. To increase the dynamic range of the assay while maintaining the simplicity, the input material was changed from a 1:1 ratio of the two strands to approximately a 1:100 ratio where the enrichment method would attempt to enrich the minority of sequences present. To evaluate the effectiveness of cycling the melt-anneal-extension-cut steps, we evaluated the fold enrichment using both one and three cycles. It was found that a 124-fold enrichment was achieved when targeting the amp strand (Fig. 3.4A, Fig. S22) while a 71-fold enrichment was achieved when targeting the hygro strand (Fig. 3.4B). When examining the effect of one vs three cycles, it appears that the first cycle contributes most to the fold enrichment, but subsequent cycling does significantly improve fold enrichment.

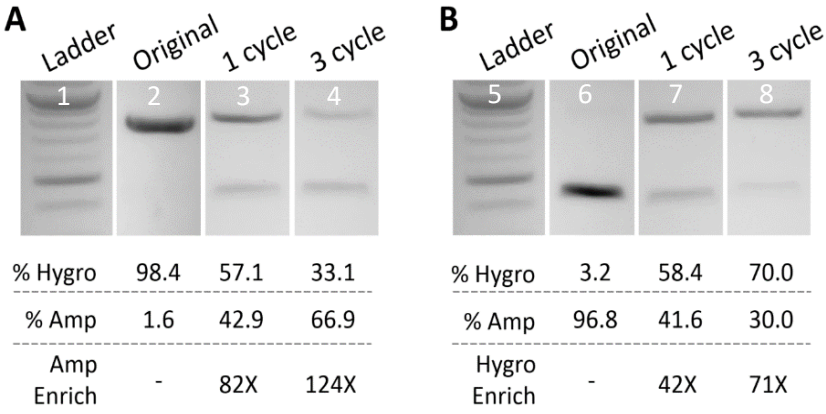


Fig. 3.4: Determining fold enrichment achieved by CleavR. The same Amp and Hygro strands from Fig. 3.3 are used here, but original mixtures are instead made in ratios of approximately 1:100 (Lane 2) or 100:1 (Lane 6) to increase dynamic range of measurements. Fold enrichment was found to increase as more cycles of CleavR were performed; (A) Hygro capture: 1 cycle with 82X (Lane 3), 3 cycle with 124X (Lane 4); (B) Amp capture: 1 cycle with 42X (Lane 7), 3 cycle with 71X (Lane 8).

3.5: Application to the immune repertoire

While validation experiments provided encouraging results, they are not indicative of a true sequencing library which consists of much higher diversity both in sequence and molecular

length. To test the enrichment method on a realistic library, we applied it to a SPLiT-seq library on T-cells with the goal of capturing the cells' receptor information.

T-cell receptors (TCRs) are made highly diverse through genetic recombination. This diversity is a critical attribute to our immune system's ability to identify and eliminate harmful population of cells. There are two genes corresponding to the TCR's beta and alpha chain that provide its diversity: TRB and TRA. Each of these genes eventually have only one variable (V),

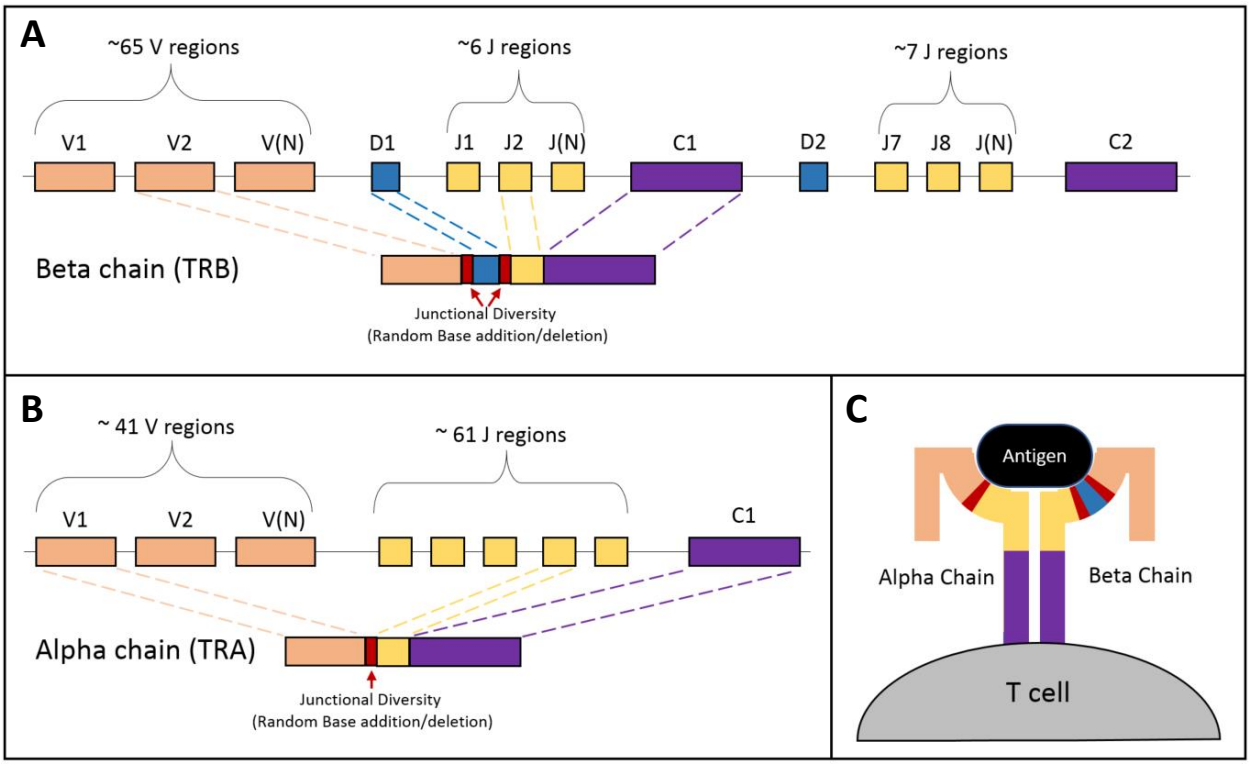


Fig 3.5: V(D)J recombination overview. (A) TRB, the gene responsible for the beta chain on the TCR, undergoes genetic recombination in T cells. The TRB gene originally has about 65 V regions, 2 D regions, about 13 J regions, and 2 constant regions. Randomly, one of each of these segments becomes somatically recombined to form the TRB gene. At each junction (V-D and D-J), additional base deletion or additional may occur to further increase the diversity – a phenomenon commonly referred to as junctional diversity. (B) Similar somatic recombination occurs in TRA, the gene responsible for the alpha chain of the TCR. While no D segment exists, enormous diversity is still generated by recombining one of about 41 V regions and one of the about 61 J regions with junctional diversity. The TRA gene has only 1 constant region. (C) Protein level representation of how the TRA and TRB genes affect TCR receptor binding to antigens, which are presented to the TCR by antigen presenting cells. The highly diverse regions determined by V(D)J recombination dictate which TCRs uniquely bind with high specificity to which antigens.

diversity (D, only for TRB), joining (J) and constant region once genetic recombination is complete. For each of these V, D, and J regions, there are many possible segments that can be chosen during genetic recombination (Fig. 3.5A). For instance, the TRA gene will consist of one of about 41 different possible variable segments and one of about 61 possible joining segments (Fig. 3.5B). A process called junctional diversity occurs during recombination where additional molecular machinery adds or subtracts bases randomly at the junctions of these segments. The high sequence diversity in these genes lead to a highly diverse TCR at the protein level (Fig. 3.5C) whereby it can detect antigens with high specificity. The sequence from both TRB and TRA genes are needed to realize the specificity of a TCR, however, TRB is located on chromosome 7 while TRA is located on chromosome 14. Because these genes are located in very different sections of the genome, single cell sequencing methods are needed to pair the alpha and beta chain sequences for T-cells.

A major initiative across both industry and academic groups has been to profile the sequence diversity of T cell populations in healthy and diseased people. Linking specific TCRs to antigens would provide a powerful resource for the development of cell-based treatments like adoptive T cell transfer and chimeric antigen receptor (CAR) therapies that rely on T-cells' ability to recognize antigens linked to cancerous cells. Furthermore, it is hypothesized that forming a TCR-Antigen map would allow individualized therapies based on every person's unique TCR sequences (and therefore the exposed antigens) in their body (*66, 67*).

To test the method for enrichment of V(D)J information from a SPLiT-seq library, we must confirm that TRA and TRB genes are present in the library. It must also be confirmed that the captured sequences from these genes span sections that provide V(D)J information. Therefore, the first step was to measure SPLiT-seq's ability to capture the complementarity determining region

3 (CDR3) on the TRA and TRB genes from each cell. The CDR3 region would supply enough information to elucidate the specific V, D, and J regions for the T cell receptor chain.

Modifications were made to the SPLIT-seq protocol to capture the CDR3 region of the TRA and TRB genes. The CDR3 region is located in the middle of these transcripts. Using poly(dT) reverse transcription primers would select for the 3' end of the transcript, outputting reads located in constant region of the transcript. Random hexamer reverse transcription primers have ability to randomly prime in any location on the transcript. The standard SPLiT-seq protocol uses half poly(dT) and half random hexamer primers. Here, the poly(dT) primers were removed and random hexamer primer concentration was doubled, providing a higher chance of capturing the CDR3 region of the TRA and TRB transcripts.

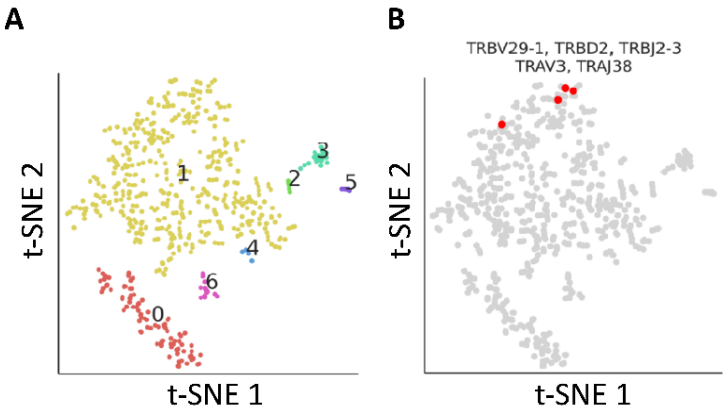


Fig. 3.6: Pilot experiment of SPLiT-seq for combined scRNA/V(D)J-seq capability. (A) *Ex vivo* expanded T-cells were sequened using SPLiT-seq and clustered using t-SNE. (B) With the same cell population that was clustered based on RNA expresion profile, specific V(D)J information for both the beta and alpha chain of the T cell receptor can be extracted. Red points highlight 4 cells of the same clonotype that cluster near each other with respect to all T cells sequenced. V, D, and J regions are provided for the beta chain as well as V and J regions for the alpha chain for this particular clonotype.

Ex vivo expanded T-cells were processed using SPLiT-seq and the resulting library was sequenced (Fig. 3.6A). For each cell, reads containing a CDR3 region were sorted and aligned (Fig. 3.6B). The TRB CDR3 region was detected in 40% of cells while the TRA CDR3 region was detected in 34% cells. However, only 0.04% of the raw sequencing reads mapped back to a TRA or TRB CDR3 region. This is not a surprise, as TRA and TRB genes are known to exhibit very

low expression in T cells. Nevertheless, this means that even if the CDR3 region from both chains of all cells was captured during SPLiT-seq, very deep sequencing would be required to read the information because only 1 out of every ~2,500 sequencing reads would have CDR3 information. With knowledge that molecules containing V(D)J information were present but at very low frequency, we moved on to applying CleavR to these molecules.

Only three captures primers are needed to enrich for V(D)J regions using CleavR (Fig. 3.7). These capture primers target highly upstream regions on the three possible constant regions (TRBC1, TRBC2, TRAC). As the variable CDR3 region is located roughly 30-50 nucleotides upstream (dependent on J region) from the constant region, molecules that contain an upstream region in the constant region have a very high probability of containing a CDR3 region.

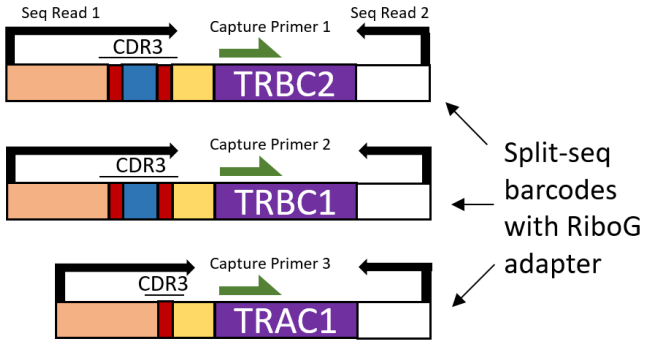


Fig. 3.7: Schematic of CleavR capture of CDR3 regions. Green arrows resemble location of capture primers on each of the three constant regions. Black arrows resemble the areas that are sequenced, with the CDR3 region on read 1 and SPLiT-seq barcodes on read 2. Notably, the CleavR capture primer binding sites are not located on the highly variable CDR3 region.

When CleavR was applied to the same SPLiT-seq library that originally yielded just 0.04% reads mapping to CDR3 regions, a ~20-fold enrichment observed with 0.77% of reads now mapping to CDR3 regions. Despite this CleavR-SPLiT-seq library being sequenced roughly 3 times lower than the original SPLiT-seq library, the TRB CDR3 region was now detectable in 57% of all cells with a TRA CDR3 region being detected in 47% of all cells. Comparing this to the original unenriched data, CleavR was able to detect roughly 40% more of these CDR3 regions with substantially fewer sequencing reads.

There are a few reasons that could explain the decrease in fold enrichment in this assay when compared to the preliminary experiments using Amp and Hygro strands. First, the sequence diversity in RNA-seq libraries would allow for more opportunities for either non-specific extension, or non-specific hybridization to the universal adapter region. Either of these scenarios would lead to protection of strands that should be depleted. Furthermore, capture primers used in this assay could be further optimized for binding efficiency as any target strand that fails to undergo extension will be depleted.

3.6 Conclusions

CleavR is a molecular target enrichment method orthogonal to any existing technology that combines the specificity of PCR without the downfalls of exponential amplification in multiplex PCR assays. CleavR requires only a few hours to perform with less than an hour of hands-on time, making it easy to add to the end of a library preparation method (i.e. SPLiT-seq). A unique advantage of CleavR is that a highly diverse region can be enriched for based on a known constant sequence downstream of the region, as demonstrated in CleavR's capture of CDR3 regions.

While 124-fold enrichment in a simple library and 20-fold enrichment in complex library were achieved with CleavR, this method has not yet been fully optimized. If strategies for improving the specificity of extension and efficiency of the T1 Rnase enzyme were implemented, it is possible this method could yield many hundreds to even thousands-fold enrichment of molecules.

Here CleavR was applied to a scRNA-seq library, however, this same strategy could easily be realized in other types of sequencing libraries. For instance, many researchers are sequencing thousands of exomes from individuals, which account for just ~1.5% of the human genome.

Furthermore, some diagnoses require detection of just a few mutations scattered throughout the genome, but sequencing of the whole genome would render the diagnostic too expensive. Moving forward, I will work to further optimize the assay to increase fold enrichment while also applying the method to different types of sequencing libraries to expand capabilities.

Chapter 4. Massively Parallel Reporter Assays for Splicing Detection

4.1: Motivation

Splicing is a process where complex molecular machinery detect intron-exon boundaries at the pre-mRNA level to include exons and exclude introns from the mRNA transcript for subsequent translation. Alternative splicing greatly enhances the diversity of proteins that the human body can produce by selectively altering the location of some exon-intron boundaries through splicing, creating two or more isoforms. It is widely accepted that about 95% of multi-exon human genes are alternatively spliced (*68*). Genetic mutations have been shown to cause aberrant splicing by inducing exon inclusion, exon skipping, or altering the normal isoform ratios. A number of genetic disorders such as Spinal Muscular Atrophy (SMA) (*69*), Alzheimer's Disease (*70*), Huntington's Disease (*71*) and many more have been linked to improper isoform ratios, stemming from the lack of regulation at pre-mRNA splicing level.

A majority of splicing mutations are single nucleotide polymorphisms (SNPs), where a single change in a base impacts splicing regulatory elements. SNPs present in regulatory splice sites such as the 5' splice donor, branchpoint, and 3' splice acceptor sites are known to impact normal isoform ratios (*72*). However, there are a number of poorly understood sites such as intronic splicing enhancers (ISE), exonic splicing enhancers (ESE), intronic splicing silencers (ISS), and exonic splicing silencers (ESS) that can preferentially increase or decrease the likelihood of a splicing event, adding a level of complexity to the regulation (*73*). Furthermore, SNPs can introduce cryptic splice sites, where a section of mRNA that previously did have a role in splicing regulation is capable of altering the isoform ratio (*73, 74*). For all of these reasons, it is poorly understood which mutations in exons will lead to aberrant splicing.

A computational model developed in the Seelig Lab called hexamer additive linear (HAL) can predict which mutations may lead to isoform ratio shifts as a result of exon skipping (*74*). While HAL is capable of computationally filtering through millions of variants and outputting splicing behavior with higher accuracy than any other model, a high-throughput experimental approach must be developed to determine any inherent errors in the splicing predictions.

Typical splicing experiments to determine a single exon's isoform ratio, also known as the percent spliced in (PSI), involve use of a minigene reporter assay. This assay includes a plasmid that contains an exon of interest (EOI) and the EOI's respective flanking intronic sequences. Most studies examine the wildtype exon and variant of the exon to measure the change in percent spliced in, or DPSI. An individual plasmid must be created for the wildtype and each variant of the EOI that is desired to be tested. For each variant, each plasmid must be assembled and cloned through bacterial transformation. Transfection into mammalian cells followed by RNA extraction, RT-PCR, and an analytical gel must be performed to understand the appropriate isoform ratio (*75–77*). While this process is tolerable when studying a few variants of a single EOI, it is impractical to use this method to study hundreds of thousands of variants over multiple EOIs.

Here we discuss the development of a universal minigene reporter assay to ultimately perform hundreds of thousands of minigene splicing experiments in a comparable amount of time as a single traditional splicing experiment. The goal of this study is to build a massively parallel reporter assay (MPRA) is to study the effect of biologically relevant SNPs with regard to splicing. To do this, we will examine the design constraints these data may reveal the most targetable regions in the pre-mRNA transcript that will result in the discovery of new optimized RNA therapeutics for currently uncured genetic disorders.

4.2: Experimental approach

Recent advances in oligonucleotide synthesis technologies allow for production of pools consisting of hundreds of thousands of unique oligonucleotides of up to 230 nucleotides (nt) long. To enable the testing of many unique sequenced in parallel, a universal plasmid backbone will be used instead of creating an individual plasmid for each variant. This backbone is capable of incorporating each of these unique oligonucleotide sequences that will ultimately allow testing of all variants in a single experiment. The overall design on the MPRA consists of two main components: unique inserts and a universal plasmid backbone.

Inserts*:*

The 230nt long oligonucleotides will act as the unique inserts. They will consist of five different domains: a 30nt 5' Gibson overlap region, a 50nt 5' intronic region native to the EOI, the EOI itself, a 20nt 3' intronic region native to the EOI, and a 30nt 3' Gibson overlap region (Fig. 4.1A). The first and last 30nts of each insert (Gibson overlap regions) will have matching sequences to the ends of the plasmid backbone, which will allow for a Gibson assembly reaction to incorporate each unique insert into the universal backbone. The flanking intronic sequences

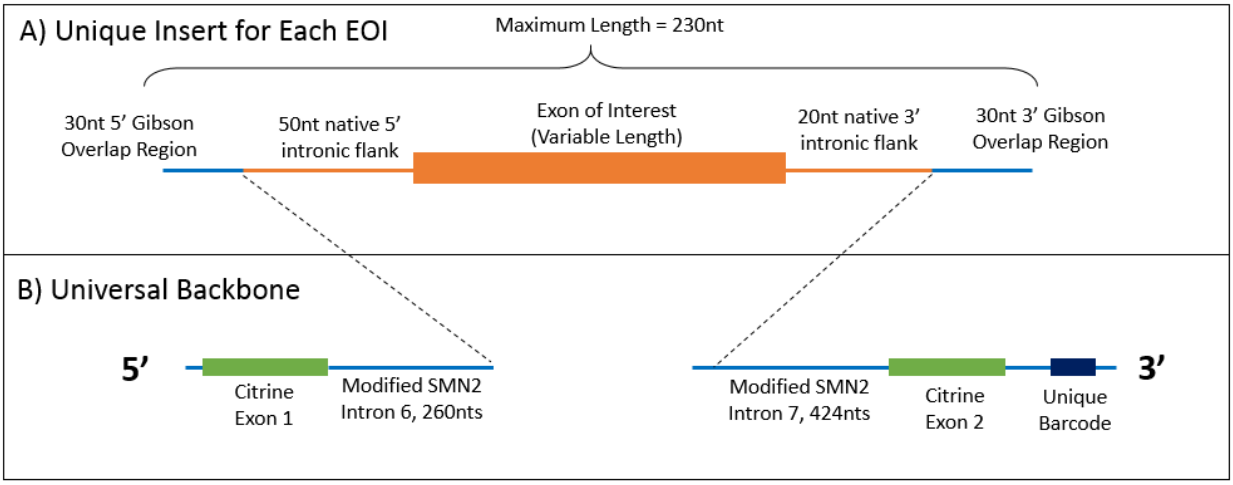


Figure 4.1: Design of the plasmid for the massively parallel reporter assay. (A) Unique oligonucleotides will be created containing an EOI and the native flanking intronic regions, in addition to a Gibson overlap region on either side to allow for assembly to the universal backbone. (B) Section of the universal backbone that will integrate with the unique inserts.

native to the EOI are included into the insert to ensure that native branchpoint, 5' splice donor, and 3' splice acceptor sites are retained. The chosen lengths of flanking intronic sequences were selected to ensure the inclusion of all necessary regulatory splice sites (*72, 78*).

Backbone:

The functional region of the universal backbone will contain five different domains: exon 1 of the citrine gene, a 260nt long region modified from SMN2 intron 6, a 424 nt long region modified from SMN2 intron 7, exon 2 of the citrine gene, and lastly a randomized sequence that will act as a plasmid barcode (Fig. 4.1B). While the native 50nt 5' and 20nt 3' native intronic regions in the insert portion are needed to maintain their native regulatory splice sites, the additional intronic sequences deriving from SMN2 intron 6 and 7 is included to ensure the exon can be spatially recognized by splicing machinery, as native exons are typically spread out by large intronic regions. For the SMN2 intron 6 portion of the backbone, 150nts of the 5' side of the intron combined with 110nts of the 3' side starting at the -50nt position. For the SMN2 intron 7 portion of the backbone, the entire intron except for the first 20nts of the 5' side will be included, equating to a total of 424nts. The randomized sequence acting as the plasmid barcode is located in the 3' UTR of the citrine gene and will be used to relate pre-spliced and post-spliced plasmids during sequencing.

4.3: Results

It is imperative that the universal backbone used for the MPRA assay simulates splicing behavior as a result of the sequence in unique inserts rather than artifacts of the assay design. There may be hidden motifs in the universal backbone that artificially upregulate or downregulate exon skipping independent of whatever sequence exists in the unique insert, which would render the assay useless. To ensure the universal backbone didn't induce any bias, a series of traditional

minigene reporter experiments were performed. Wild types and variants of well-studied exons were used as unique inserts and cloned using the universal backbone. The exons tested were SMN2 exon 7 (54nts long), MAPT exon 10 (93nts long), and DMD exon 29 (150nts long). All of these exons have been significantly studied and have documented variants that have corresponding isoform ratios(*79–82*). Variants leading both to high exon inclusion and high exon skipping were tested to ensure the universal backbone inflicts no bias on either event, but rather simulates splicing as a direct result of the mutation present in the variant. For each of these three exons, the wildtype

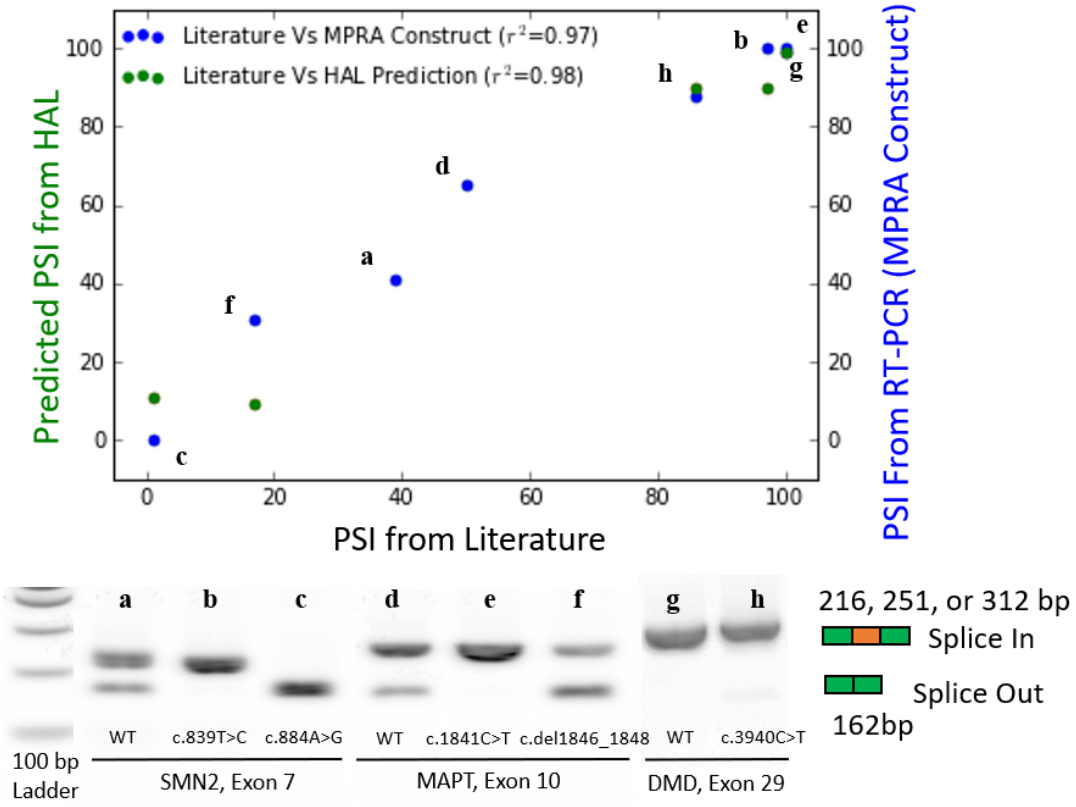


Figure 4.2: Minigene splicing experiments using the universal backbone. The wildtype (WT) sequence of SMN2 exon 7, MAPT exon 10, and DMD exon 29 with corresponding variants that have been well studied in literature were inserted into the universal backbone. The PSI for each minigene experiment was calculated by measuring band intensities through ImageJ. This experimental PSI was then compared to values found in the literature (blue dots on the plot) where a high correlation was found (r^2 = 0.97). These values were also compared to predictions of HAL, where a high correlation was also observed (r^2 =0.98). Note that HAL cannot make predictions on the wildtype PSI as the wildtype PSI is a user defined input for the model (and would therefore only report what the user has defined the WT PSI to be). HAL predictions were made based on WT PSI from literature, and not those found by WT sequence in the universal backbone.

and different variants documented from literature were cloned into the universal backbone. Traditional minigene reporter experiments were performed for each plasmid to assess the PSI for wildtypes and DPSI for mutants. This was quantified by RT-PCR from the flanking citrine exon 1 the other flanking exon, citrine exon 2, such that the PCR amplicon spans the exon junction. Therefore, if an exon is included the amplicon will be longer than an amplicon where an exon was spliced out. These amplicons can be run on a gel to determine the PSI and DPSI ratios. When these three exons and their corresponding variants where measured within the context of the universal backbone plasmid, it was found that their PSI/DPSI was highly correlated to both values found in literature and the predicted values outputted by HAL (Fig. 4.2).

With validation that the MPRA design has no substantial effect on native splicing mechanisms, the next steps are to select exons and variants to test. Variants to test will be determined by extracting previously known pathogenic mutations or variants of unknown significance that are predicted to cause significant splicing events. Some databases such as ClinVar have documented variants that are known to cause exon skipping, which will be used in this screen to serve as a broader validation. ClinVar, in addition to the Exome Aggregation Consortium (ExAC) database, will also be searched for variants present in alternatively spliced exons. We will use HAL to predict how these variants will affect the PSI of the exon. Variants predicted to cause a high DPSI will be selected for the experimental screen.

4.4: Predicting splice switching oligonucleotide performance

Splice switching oligonucleotides (SSO), also known as antisense oligonucleotides, are an FDA-approved therapy that targets the pre-mRNA transcript to locally modify splicing behavior of a single exon. SSOs are small oligonucleotides (about 14-20mer) designed to bind to an exact site on the pre-mRNA transcript to induce exon skipping or inclusion. While only two SSOs are

commercially available, there are a few currently in human clinical trials (*70*, *83*). One of the largest barriers to SSO drug discovery is determining what sequence in the pre-mRNA should be targeted. While attempts have been made to formulate a high throughout method for SSO selection, none have been successful (*84*, *85*). These methods involve using ESE finding software, which searches a user defined exonic sequence for motifs that correspond to known splicing enhancer sites as reported from various literature sources. The idea here is if splicing enhancer sites can be blocked, there may exist a higher probability that the exon will be skipped. While this approach may sound effective, the programs fail to find many ESE motifs, may detect an ESE motif where no such site exists, and fail to accurately rank the strength of an ESE (*86*). These are some of the reasons why this method often leads to exhaustive trial and error experiments to find SSOs (*87*, *88*). These experiments result in slow, expensive, and sub-optimal SSO discovery, revealing the need for a high throughput method of determining optimized SSO target sequences.

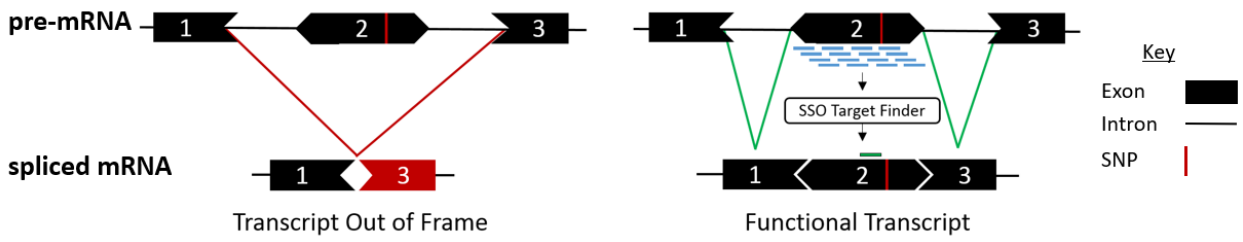


Fig. 4.3: Schematic demonstrating SSO target finder. A SNP present in the pre-mRNA in exon 2 causes aberrant splicing and leads to exon skipping, leading to an out of frame transcript and loss of function. SSO target finder can be implemented to screen for all possible SSOs along exon 2. The output is a single SSO that is most efficient at promoting exon inclusion, therefore restoring the spliced mRNA to a functional transcript.

HAL demonstrated best-in-class performance at predicting exon skipping as a consequence of a SNP. Here we extend the capabilities of HAL in attempts to provide a ranked list of optimal SSO binding sites for exon inclusion or exclusion - a SSO target finder (Fig. 4.3). This was done by first building significance scores for each nucleotide present in a given exon of interest (i.e. exon where increased inclusion/exclusion is desired). These scores are created by averaging the HAL DPSI scores for all 4 possible SNPs that can occur at a given location in an exon (3 other bases and a deletion). An exon-spanning heat map can then be generated using these scores where 14-20 base regions that have sustained high or low DPSI significance scores could indicate sites along the pre-mRNA transcript that may be particularly favorable for SSOs. Because these scores are essentially predicting the

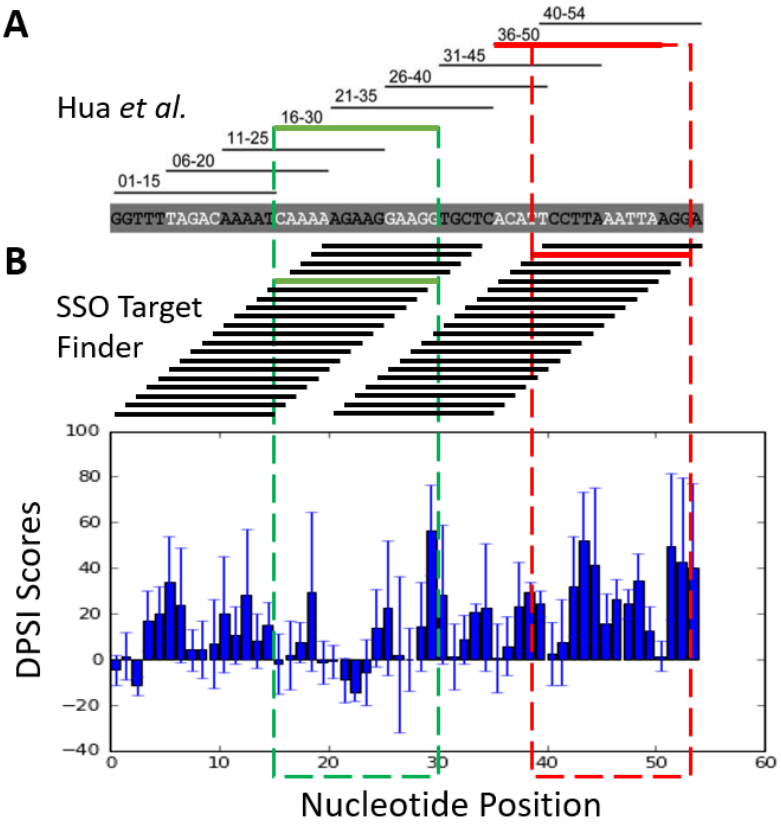


Fig. 4.4: Comparing SSO target finder to literature. (A) Image section adapted from Hua *et al* where they tested nine SSOs spanning exon 7 of the SMN2 gene for exon inclusion/exclusion. SSO spanning bases 16-30 was found to promote highest exon exclusion while SSO spanning based 36-50 was found to promote highest exon inclusion. (B) SSO target finder computationally screened all possible 15mer SSOs spanning the exon. Aggregate DPSI scores, calculated using individual values for each base from the bar plot, were used to predict each SSOs likelihood of promoting exon inclusion/exclusion. SSO target finder predicted SSO spanning bases 16-30 to promote highest exon exclusion, identical to results found in Hua *et al*. Further, SSO spanning bases 39-53 was predicted to promote highest inclusion, just 3 bases off from the screen done by Hua *et al.*

inclusion/exclusion due to changing part of an exon's sequence, the hypothesis is that a 14-20mer region with the highest sustained DPSI could resemble a section of the exon particularly important to preventing exon inclusion. This would highlight a vulnerable site in the exon, where a large shift towards exon inclusion could be induced if an SSO bound there. The inverse could also be true, where regions with lowest sustained DPSI scores could resemble an SSO target site if exon exclusion was desired.

Validating a computational tool for SSO efficacy is difficult mostly due to the sparsity in experimental data. Very few published studies have demonstrated a comprehensive testing of multiple AOs across the same exon. Of the studies that have done these types of large-scale measurements, many choose not to release the raw data indicating the DPSI for each SSO tested. However, some data does exist on the SMN2 exon 7, where Hua *et al* screened many different SSOs on this same exon and reported their DPSIs (*87*) (Fig. 4.4A). When the HAL-inspired SSO target finder attempted to computationally determine the SSO that would yield the highest exon exclusion, the predicted SSO was identical to the best performing exon exclusion SSO as reported Hua *et al* (Fig. 4.4B, Fig. S23). Furthermore, the SSO predicted to promote highest exon inclusion was just 3 bases off from the best performing exon inclusion SSO as reported in Hua *et al*. Notably, this predicted SSO was not experimentally tested in the Hua *et al* study. While promising, significantly more data on SSO screens are needed to properly validate the SSO target finder. Looking forward, we will look to collaboration with SSO drug companies who presumably have large datasets on these screens for further validation.

4.4: Conclusions

There is much to learn about how mechanisms of splicing impact disease. The ability to screen thousands of variants in parallel with MPRAs will accelerate our understanding of this

process and may lead to characterization of variants that currently have unknown significance. Here, I discussed how to design and properly validate an MPRA to measure exon skipping. I then took computational models originally built to assess effects of SNPs on exon skipping and extended it to predict optimal SSOs that could promote either exon inclusion or exclusion.

Chapter 5: Towards Translation of Developed Technologies

5.1: Website

Once SPLiT-seq was published, we received a great deal of interest from others who wanted to try it out. To facilitate adoption of the method, we created a website (https://sites.google.com/uw.edu/splitseq) where we posted the protocol along with other materials such as frequently asked questions and sequence information for all oligonucleotides used. To date, over 7,000 unique users from around the world have visited the site (Fig. 5.1). While a handful of labs have communicated to us their success with the protocol, it is difficult to know how many labs have been successful in executing the protocol as we suspect many simply haven't felt the need to tell us. I am so appreciative of all the groups who have reached out with suggestions and even more so to those who have spent their own time getting it to work – it has been the ultimate motivator.

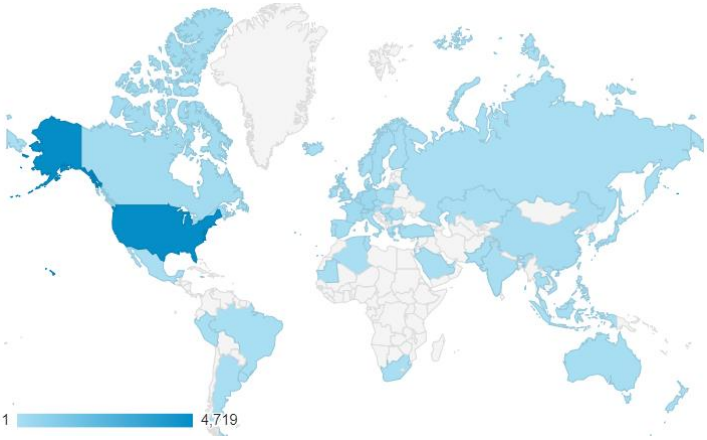


Fig. 5.1: Geographic overview of users visiting the SPLiT-seq website

5.2: Kits

While we tried our best to make the online protocol as clear as possible, it was to our disappointment that many users were still having difficulty to get SPLiT-seq up and running. When talking with users who were unable to succeed, we realized there were a few common issues. Often time reagents were being substituted out with different ones that were readily available in their lab. The need to make multiple master mixes containing many different reagents often led to simple pipetting mistakes. Additionally, many steps were being slightly modified that we found were

having an impact on the final results. Since the protocol can take 2-3 days, there is ample opportunity for just one pipetting mistake, or for a step in the protocol to be performed slightly differently. We thought that the creation of a kit would help standardize reagents, minimize the pipetting steps for users, and streamline the overall protocol to prevent unwanted deviations.

With some funding from the Washington Research Foundation (WRF), we were able to kit together the reagents needed to execute SPLiT-seq. We distributed these kits throughout labs at various institutions in the Seattle area including the Fred Hutch Cancer Research Center, the Allen Institute, and Seattle Children's Research Institute (Fig. 5.2). In addition to each lab running their own samples, we also provided out own internal control sample consisting of HEK293 and NIH/3T3 cells mixed at even proportions, similar to the species mixing experiments described in chapter one. The goal of this control sample was to understand if results could be reproduced across multiple users, as it would be impossible to properly compare different experiments that had very different sample types.

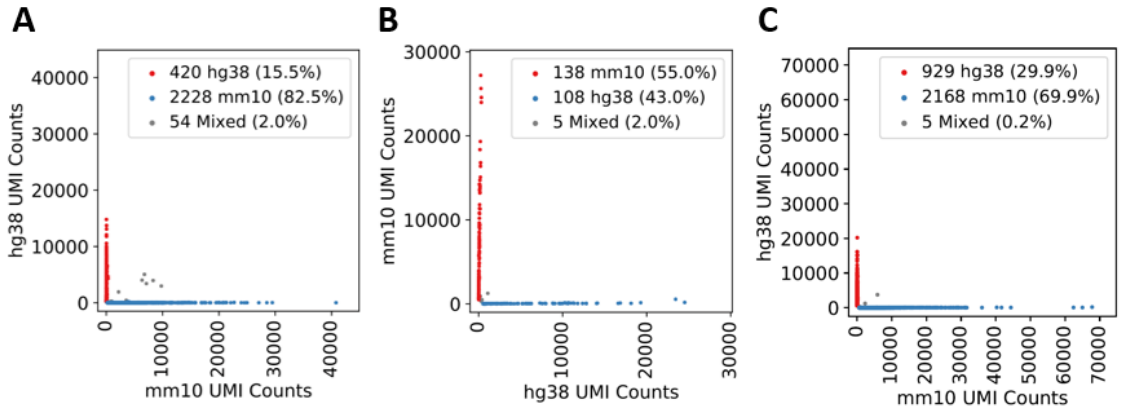


Fig. 5.2: Species-mixing experiment results from external kits. Three different groups from (A) the Fred Hutch Cancer Research Center, (B) Seattle Children's Research Institute, and (C) the Allen Institute executed a SPLiT-seq kit using our control cells along with their samples.

Comparing results of the control cells across multiple kit tests, we found that the doublet rate was consistently quite low at rates lower than 5% while gene detection remained consistent

with our in-house results. Moving forward, we will look to expand availability of these kits for other researchers wanting to perform scRNA-seq.

Supplementary Text

SPLiT-seq Barcode Collisions

Barcode collisions result from two scenarios: a mouse and human cell form a physical doublet and remain stuck to each other for the entirety of split-pool barcoding experiment, or a mouse and human cell happen to be barcoded with the same combination of barcodes by chance. The first issue can in principle be addressed by FACS sorting cells before barcoding, the second by increasing the number of barcode combinations (either by adding additional barcoding rounds or by switching to 384-well plates for any or all of the barcoding rounds).

Cluster Identification of Mouse Brain and Spinal Cord Cell Types

The 54 neuronal clusters deriving from the brain were characterized using a number of known markers. Three neuronal clusters (clusters 1-3) were determined to be mitral/tufted cells (*Tbx21+*, *Deptor+*) (*89*, *90*), types of projection neurons specific to the olfactory bulb. Consistent with previous work (*91*), medium spiny neurons (cluster 4) expressed markers specific to the striatum (*Rarb*, *Drd2*). More than 22,000 cortical pyramidal neurons clustered into 14 types (clusters 5-18), with nearly all expressing the pan-excitatory cortical marker *Satb2* (*92*). Using known markers (*10*, *11*), we were able to further assign most cortical types to specific layers (Fig. 3C).

We assigned cluster 19 to the rostral midbrain based on unique expression of *Tfap2d*, a gene known to be required in midbrain development (*93*). We identified one excitatory (cluster 20, *Slc17a6*$^{+}$) and one inhibitory neuron (cluster 21, *Gad1/2*$^{+}$) type originating from the thalamus (both *Fign+*). Eight different types of neurons expressed markers specific to the cerebellum (clusters 22-29). Expression of *Pax2* was found in two inhibitory interneuron types from the medulla (clusters 30 and 31), consistent with ISH data (*31*). Nigral dopaminergic neurons (cluster

32) were identified based on specific expression of the dopamine transporter *Slc6a3* (*35*), whose expression is restricted to the substantia nigra of the basal ganglia (*31*). Nine types of pyramidal cells and granular cells were inferred to have originated from the hippocampus (clusters 33-41). Excitatory neurons from the spinal cord (clusters 42 and 43) were marked by specific expression of *Pde11a*. We found eight types of migrating GABAergic interneurons (clusters 44-51, *Gad1/2+*), based on expression of members of the *Dlx* family (*94*). Cajal-Retzius cells (cluster 52) expressed *Trp73*, which is expressed in the P4 hippocampus and the marginal zone of the cortex, confirming their known distribution (*95*). Clusters 53 and 54 contained pan-neuronal markers, but could not be assigned to a specific cell type. After inspection of sample distribution for the unresolved clusters (Fig. S19), it was found that these clusters represented neurons originating from the spinal cord. Re-embedding of these cells resulted in substantial increase in resolution (Fig. 5).

There were 19 non-neuronal clusters composed of 27,096 individual transcriptomes. We identified 6 types of oligodendrocytes (clusters 55-60) and one OPC cluster (cluster 61), which together formed a lineage (Fig. 2D-E, Fig. S7, Fig. S8). Immune cells expressed the pan-immune marker *Dock2*, but *Ly86* expression was restricted to microglia (cluster 63), whereas *Mrc1* expression occurred only in macrophages (cluster 62, Fig. S5) (*26*, *27*). Both types of vascular cells expressed *Rgs5*, with endothelial cells (cluster 64) marked by distinct expression of *Flt1* and *Kdr* (*10*) and smooth muscle cells (cluster 65) marked by expression of *Abcc9* and *Pdgfrb* (Fig. S5) (*96*). We identified two vascular and leptomeningeal cell (VLMC) subtypes (clusters 67 and 66), both expressing previously characterized markers *Col1a1* and *Pdgfra* (*12*). Cluster 67 VLMCs expressed *Slc47a1* and *Slc47a2*, while cluster 66 VLMCs specifically expressed *Slc6a13* (Fig. S5).

Astrocytes were the most abundant non-neuronal cell type, accounting for 50% of all non-neuronal nuclei (n=13,481). Among the four astrocyte types (all *Aldh1l1+*), only Bergman glia

(cluster 71) expressed *Gria1* (Fig. 2C) (*22*). Cluster 70 astrocytes—found only in the spinal cord—expressed *Gfap* highly, while cluster 69 astrocytes—found almost exclusively in the brain—expressed *Prdm16*. Cluster 68 astrocytes—found in both the brain and spinal cord—were defined by specific expression of *Slc7a10*. Ependymal cells (cluster 72) uniquely expressed many previously characterized markers (*11*) such as *Foxj1* and *Dnah1/2/5/9/10/11*. Olfactory ensheathing cells (OEC, cluster 73), a type of Schwann cell specific to the olfactory bulb (*26*), were identified by unique expression of *Mybpc1*, a gene expressed specifically in the outer layers of the olfactory bulb (*27*).

Among the 30 neuronal clusters found in the spinal cord, 28 neuronal types were identified using markers from previous literature. Clusters highly expressing *Gad1/2* were determined to be GABAergic neurons. Subsets of these GABAergic neurons included glycinergic neurons, marked by *Slc6a5*, and cerebrospinal fluid-contacting neurons, marked by *Pkd2l1* and *Pkd1l2* (*55*). Clusters expressing *Slc17a6* (VGlut1) were identified as glutamatergic neurons. A subset of these glutamatergic neurons also expressed *Slc17a8* (VGlut3). The two cholinergic motor neurons (alpha and gamma) were identified with *Chat* expression (*56*).

Cost Analysis of SPLiT-seq

An itemized cost analysis of SPLiT-seq was conducted (Table S11) for an experiment using two sublibraries (884,000 barcode combinations: 48 x 96 x 96 x 2), which makes it possible to sequence 44,000 cells with an expected 5% barcode collision rate at a cost of $0.02 per cell. If six sublibraries are used (2.65 million barcode combinations: 48 x 96 x 96 x 6), more than 132,000 cells could be processed at a cost of $0.01 per cell. Most costs derive from reverse transcription and ligation enzymes. The price per cell drops dramatically with scale because experimental costs do not increase linearly with cell numbers. Adding additional sublibraries does marginally increase

costs, largely due to the use of more template switch primer, reverse transcriptase, polymerase, and Nextera reagents.

Costs to get data on scRNA-seq from hundreds of thousands of cells using commercial methods

A conservative cost estimate to acquire a single cell RNA-seq library on the current most common scRNA-seq platform, 10X Genomics Chromium system, is $0.18 per cell sequenced (*97*, *98*) (Fig. 3.1A). SPLiT-seq is now capable of acquiring similar quality data for a conservative estimate of about $0.02 per cell. However, this leads to no substantial effect on sequencing cost (Fig. 3.1B). While the amount of sequencing required for each experiment can vary widely, a general base amount of sequencing lands around 30,000 raw sequencing reads per cell. To capture information from rare transcripts, sequencing of 75,000+ raw sequencing reads per cell may be required. Using rates from Illumina's most scalable Novaseq instrument, this would breakdown to base sequencing costing around $0.12 per cell, with costs going up to $0.3 per cell when looking for more rare transcripts.

Wet Lab Methods

Cell Culture

HEK293 and Hela-S3 cells were cultured in DMEM + 10% FBS, while NIH/3T3 cells were cultured in DMEM + 10% FCS. Cells were rinsed twice with 1x PBS, then detached by incubating 2-5 min at room temperature with 1ml of TrypLE. Once cells were detached, they were added to 2mL of media with 10% FBS. In mouse-human species mixing experiments, cells were combined at the desired concentrations at this step.

Fixation

Cells were first centrifuged for 3 min at 500g at 4˚C. The pellet was resuspended in 1mL of cold PBS-RI, 1x PBS + 0.05U/μL RNase Inhibitor (Enzymatics). The cells were then passed through a 40 μm strainer into a 15 mL falcon tube. 3 mL of cold 1.33% formaldehyde solution (in 1x PBS) was then added to 1 mL of cells. Cells were fixed for 10 min before adding 160 μL of 5% Triton X-100. Cells were then permeabilized for 3 min and centrifuged at 500g for 3 min at 4˚C. Cells were resuspended in 500uL of PBS-RI before adding 500 μL of cold 100 mM Tris-HCL pH 8. In order to make the cells easier to pellet, 20 μL of 5% Triton-X100 was added. Then, cells were spun down at 500g for 3 min at 4˚C and resuspended in 300 μL of cold 0.5 X PBS-RI. Finally, cells were again passed through a 40 μm strainer into a new 1.5 mL tube. Cells were then counted on a hemocytometer or flow-cytometer and diluted to 1,000,000 cells/mL.

In-cell Reverse Transcription

The first round of barcoding occurs through an *in situ* reverse transcription (RT) reaction. Cells are split into up to 48 wells, each containing barcoded well-specific reverse transcription primers. Both random hexamer and anchored poly(dT)$_{15}$ barcoded RT primers were used in each

well (Table S1). For each well, we added 4 µL of 5X RT Buffer, 0.625 µL of RNase-free water, 0.125 µL RNase Inhibitor (Enzymatics), 0.25 µL µL SuperaseIn RNase Inhibitor (Ambion), 1 µL of 10 mM dNTPs each (ThermoFisher), 2 µL of 25 µM random hexamer barcoded RT primer, 2uL of 25 µM poly(dT)$_{15}$ barcoded RT primer, 2 µL of Maxima H Minus Reverse Transcriptase (ThermoFisher), and 8 µL of cells in 0.5X PBS-RI. The plate incubated in a thermocycler for 10 min at 50˚C before cycling for three times at 8˚C for 12s, 15˚C for 45s, 20˚C for 45s, 30˚C for 30s, 42˚C for 2 min, and 50˚C for 3 min, followed by a final step at 50˚C for 5 min. RT reactions are pooled back together into a 15 mL falcon tube. After adding 9.6 µL of 10% Triton X-100, cells were centrifuged for 3 min at 500g at 4˚C. The supernatant was removed and cells were resuspended in 2 mL of 1X NEB buffer 3.1 with 20 µL of Enzymatics RNase Inhibitor.

Preparing Oligonucleotides for Ligations

The second and third barcoding round consist of a ligation reaction. Each round uses a different set of 96 well barcoding plates (Table S1). Ligation rounds have a universal linker strand with partial complementarity to a second strand containing the unique well-specific barcode sequence added to each well. These strands were annealed together prior to cellular barcoding to create a DNA molecule with three distinct functional domains: a 5' overhang that is complementary to the 3' overhang present on the cDNA molecule (may originate from RT primer or previous barcoding round), a unique well-specific barcode sequence, and a 3' overhang complementary to the 5' overhang present on the DNA molecule to be subsequently ligated (Fig. S1A). For the third round barcodes, the 5' overhang also contains a unique molecular identifier (UMI), a universal PCR handle, and a biotin molecule. Linker strands and barcode strands (IDT) for the ligation rounds are added to RNase-free 96 well plates to a total volume of 10 µL/well with the following concentrations: round 2 plates contain 11 µM linker strand (BC_0215) and 12 µM

barcodes and round 3 plates contain 13 μM linker strand (BC_0060) and 14 μM barcodes. Strands for ligation barcoding rounds are annealed by heating plates to 95°C for 2 minutes and cooling down to 20°C at a rate of -0.1°C per second.

Blocking strands are complementary to the 3' overhang present on the DNA barcodes used during ligation barcoding rounds. Blocking occurs after well-specific barcodes have hybridized and were ligated to cDNA molecules, but before all cells are pooled back together. Blocking ensures that unbound DNA barcodes cannot mislabel cDNA in future barcoding rounds. 10 μL of blocking strand solution was added to each of the 96 wells after each round of hybridization and ligation of DNA barcodes. Blocking strand solutions were prepared at a concentration of 26.4 μM (BC_0216) for round 2 and 30.8 μM (BC_0066) for round 3. Blocking strands for the first two rounds were in a 2.5X T4 DNA Ligase buffer (NEB) while the third round was in a 150 mM EDTA solution (to terminate ligase activity). Blocking strands were incubated with cells for 30 min at 37°C with gentle shaking (50 rpm).

In-cell Ligations

A 2.04 mL ligation mix was made containing 1,287.5 μL of RNase-free water, 500 μL 10X T4 Ligase buffer (NEB), 100 μL T4 DNA Ligase (400 U/μL, NEB), 40 μL RNase inhibitor (40 U/μL, Enzymatics), 12.5 μL SuperaseIn RNase Inhibitor (20 U/μL, Ambion), and 100 μL of 5% Triton-X100. This ligation mix and the 2 mL of cells in 1X NEB buffer 3.1were added to a basin and mixed thoroughly to make a total of 4.04 mL.

Using a multichannel pipet, 40 μL of cells in ligation mix were added to each of the 96 wells in the first-round barcoding plate. Each well already contained 10 μL of the appropriate DNA barcodes. The round 2 barcoding plate was incubated for 30 min at 37°C with gentle shaking (50

rpm) to allow hybridization and ligation to occur before adding blocking strands. Cells from all 96 wells were passed through a 40 μM strainer and combined into a single multichannel basin, where an additional 100 μL of T4 DNA Ligase was added. Subsequent steps in round 3 were identical to round 2, except that 50 μL of pooled cells were split and added to barcodes in round 2 (total volume of 60 μL/well). This adjustment was made to account for increased total volume during each split-pool round as well as pipetting errors.

Lysis and Sublibrary Generation

After the third round of barcoding, 70 μL of 10% Triton-X100 is added to the cell solution before spinning it down for 5 min at 1000G and 4˚C. We carefully aspirated the supernatant, leaving about 30 μL to avoid removing the pellet. We then resuspended the cells in 4 mL of wash buffer (4 mL of 1X PBS, 40 μL of 10% Triton X-100 and 10 μL of SUPERase In RNase Inhibitor) and spun down for 5 min at 1000G at 4˚C. We then aspirated the supernatant and resuspended in 50 μL of PBS-RI. After counting cells, we aliquoted them into sublibraries (in 1.7 mL tubes). The number of sublibraries generated will determine how many splits are made for the fourth round of barcoding. After adding the desired number of cells to each sublibrary, we brought the volume of each to 50 μL by adding 1x PBS, then added 50 μL of 2X lysis buffer (20 mM Tris (pH 8.0), 400 mM NaCl, 100 mM EDTA (pH 8.0), and 4.4% SDS) and 10 μL of proteinase K solution (20mg/mL). We incubated cells at 55˚C for 2 hours with shaking at 200 rpm to reverse formaldehyde crosslinks. Afterwards, we froze lysates at -80°C.

Purification of cDNA

We first prepared 40 μL Dynabeads MyOne Streptavidin C1 beads (ThermoFisher) per sublibrary by washing them 3x with 800 μL of 1X B&W buffer with 0.05% Tween-20 (refer to

manufacturer's protocol for B&W buffer), before resuspending beads in 100 µL 2X B&W buffer (with 2 µL of SUPERase In Rnase Inhibitor) per sample.

To inhibit residual proteinase K activity, we added 5 µL of 100 µM PMSF to each thawed lysate and incubated at room temperature for 10 minutes. We then added 100 µL of resuspended Dynabeads MyOne Streptavidin C1 (ThermoFisher) magnetic beads to each lysate. We then allowed binding to occur for 60 min at room temperature (with agitation on a microtube foam insert). The beads were washed twice with 1X B&W buffer and once more with 10mM Tris containing 0.1% Tween-20 (with each wash including of 5 min of agitation after resuspension of beads).

Template Switch

Streptavidin beads with bound cDNA molecules were resuspended in a solution containing 44 µL of 5X Maxima RT buffer (ThermoFisher), 44 µL of 20% Ficoll PM-400 solution, 22 µL of 10 mM dNTPs each (ThermoFisher), 5.5 µL of RNase Inhibitor (Enzymatics), 11 µL of Maxima H Minus Reverse Transcriptase (ThermoFisher), and 5.5 µL of 100uM of a template switch primer (BC_0127). The template switch primer contains two ribonucleic guanines followed by a locked nucleic acid guanine at the end of the primer (Exiquon). The beads were incubated at room temperature for 30 minutes and then at 42°C for 90 minutes with gentle shaking. Read structure after this step should resemble the sequence map on Fig. 24.

PCR

After washing beads once with 10 mM Tris and 0.1% Tween-20 solution and once with water, beads were resuspended into a solution containing 110 µL of 2X Kapa HiFi HotStart Master Mix (Kapa Biosystems), 8.8 µL of 10 µM stocks of primers BC_0062 and BC_0108, and 92.4 µL

of water. PCR thermocycling was performed as follows: 95°C for 3 mins, then five cycles at 98°C for 20 seconds, 65°C for 45 seconds, 72°C for 3 minutes. After these five cycles, Dynabeads beads were removed from PCR solution and EvaGreen (Biotium) was added at a 1X concentration. Samples were again placed in a qPCR machine with the following thermocycling conditions: 95°C for 3 minutes, cycling at 98°C for 20 seconds, 65°C for 20 seconds, and then 72°C for 3 minutes, followed by a single 5 minutes at 72°C after cycling. Once the qPCR signal began to plateau, reactions were removed.

Tagmentation

PCR reactions were purified using a 0.8X ratio of SPRI beads (Kapa Pure Beads, Kapa Biosystems) and cDNA concentration was measured using a qubit. For tagmentation, a Nextera XT Library Prep Kit was used (Illumina). 600 pg of purified cDNA was diluted in water to a total volume of 5 μL. 10 μL of Nextera TD buffer and 5 μL of Amplicon Tagment enzyme were added to bring the total volume to 20 μL. After mixing by pipetting, the solution was incubated at 55°C for 5 minutes. A volume of 5 μL of neutralization buffer was added and the solution was mixed before incubation at room temperature for another 5 minutes. In this order, a 15 μL volume of Nextera PCR mix, 8 μL of water, and 1 μL of each primer (P5 primer: BC_0118, one indexed P7 primer: BC_0076-BC_0083) at a stock concentration of 10 μM was added to the mix, making a total volume of 50 μL. Using distinct, indexed PCR primers, this PCR reaction can be used to add a unique barcode to each sublibrary barcoded. PCR was then performed with the following cycling conditions: 95°C for 30 seconds, followed by 12 cycles of 95°C for 10 seconds, 55°C for 30 seconds, 72°C for 30 seconds, and 72°C for 5 minutes after the 12 cycles. 40 μL of this PCR reaction was removed and purified with a 0.7X ratio of SPRI beads to generate an Illumina-

compatible sequencing library. The read structure after this step should resemble the sequence map on Fig. S25.

Illumina Sequencing for SPLiT-seq

Libraries were sequenced on MiSeq or NextSeq systems (Illumina) using 150 nucleotide (nt) kits and paired-end sequencing. Read 1 (66 nt) covered the transcript sequences. Read 2 (94 nt) covered the UMI and UBC barcode combinations. The index read (6 nt), serving as the fourth barcode, covered the sublibrary indices introduced after tagmentation.

For CleavR-SPLiT-seq libraries targeting the CDR3 region of T-cells, 300 nucleotide (nt) kits were used with paired-end sequencing. Read 1 (216 nt) covered the transcript sequences – this sequence is longer than typical cDNA libraries to ensure coverage of the CDR3 region. Read 2 (94 nt) covered the UMI and UBC barcode combinations, while the index read (6 nt) sserved as the 4th barcode, similar to normal SPLiT-seq libraries.

Mouse Brain and Spine Nuclei Extraction

Brain and spinal cord tissue was harvested from two postnatal mouse pups (P2 and P11) that had been exsanguinated by transcardial saline perfusion. The mouse strain used was C57BL/6 x DBA/2. All animal procedures were done using protocols approved by the Institutional Animal Care and Use Committee at the University of Washington.

Nuclei extraction protocol was adapted from Krishnaswami *et al* (*23*). Briefly, a NIM1 buffer was made consisting of 250 mM sucrose, 25 mM KCl, 5 mM $MgCl_2$, and 10 mM Tris (pH=8.0). A homogenization buffer was made consisting of 4.845 mL of NIM1 buffer, 5 µL of 1 mM DTT, 50 µL of Enzymatics RNase Inibitor (40U/µL), 50 µL of SuperaseIn RNase Inhibtor (20U/µL), and 50 µL of 10% Triton-X100.

A 1 mL dounce homogenizer (Wheaton, cat. no. 357538) was used for nuclei extraction. After adding mouse brain and spinal cord tissue, 700 μL of homogenization buffer was added to the douncer. Then 5 strokes of loose pestle followed by 10-15 strokes of tight pestle were performed. Homogenization buffer was added up to a volume of 1 mL. The homogenate was filtered with a 40um strainer into 5mL Eppendorf tubes and then spun down for 4 minutes at 600g at 4˚C. After removing supernatant, the pellet was resuspended in 1 mL of 1X PBS-RI. Then 10 μL of BSA was added and solution was spun down again for 4 min at 600g at 4˚C. Nuclei were then passed through a 40 μm strainer once more before being counted.

Implementing CleavR Target Enrichment

Amp and Hygro strands were amplified from the ampicillin and hygromycin resistance genes found on a plasmid. Primers BC_0359 and BC_0360 for the Amp strand and BC_0357 and BC_0358 were used for the Hygro strand. These primers contain 3 riboG bases throughout the universal adapter. Primers used to amplify SPLiT-seq libraries were BC_0385 and BC_0386, which effectively replace primers BC_0062 and BC_0108 in the original protocol. Capture primer for the Hygro and Amp strands were BC_0301 and BC_0306, respectively. For capture of CDR3 regions from T-cells, three capture primers complimentary to the TRAC1 (BC_0391), TRBC1 (BC_0392), and TRBC2 (BC_0393) regions were used.

Once amplified DNA containing universal adapters with riboG bases is attained, a total of 5 ng of DNA is added to a mixture containing 1.6 uL of 2.5 mM dNTPs (each), 0.1 uL of hotstart taq polymerase (5U/uL, New England Biolabs), 1 uL of T1 Rnase (100U/uL, New England Biolabs), 1 uL of capture primer(s) (2uM per each primer), and 13.3 uL of water to make a total reaction volume of 20 uL. Next the reaction is subjected to a thermocycling protocol as follows: 95˚C for 30 seconds, then one or three cycles of 95˚C for 30 seconds, 55˚C for 30 seconds, 68˚C

for 2 minutes, and 37˚C for 15 minutes, with a final resting step at 4˚C forever. Once the CleavR thermocycling has finished, an SPRI cleanup was performed by using 40 uL of Kapa Pure Beads and following the manufacturer's protocol. A PCR step is then performed with primers BC_0062 and BC_0108, which contain primer binding sites for the universal adapter region. An additional SPRI cleanup is performed before proceeding with library preparation steps previously described for the SPLiT-seq protocol.

To calculate CleavR's fold enrichment, gel images from before and after enrichment of one strand over second strand are compared. Gels are scanned using a Biorad Pharos FX Molecular Imager and the intensity of gel bands resembling these two strands are quantified using ImageJ. In most cases, the combination of intensity of these two bands contributed to 100% of the lane intensity, indicating little or no background in the gel. Using each band's intensity, a ratio can be calculated from before enrichment and after enrichment.

MPRA Splicing Backbone Validation

The universal backbone for the MPRA was validated by inserting specific WT exons with known PSIs along with those same exons with some variants with known DPSIs. . Exons tested were SMN2 exon 7 (54nts long), MAPT exon 10 (93nts long), and DMD exon 29 (150nts long). To clone these exons and their corresponding variants into the universal backbone, a six or eight part Gibson assembly reaction (*99*) was performed with DNA fragments to construct the exon of interest. For SMN2 exon 7, the wildtype was constructed using oligonucleotides S_F1, S_F2, S_F3, S_F4, S_F5, and S_F6. Construction for variant c.839T>C replaced S_F2 and S_F3 with S_V1F2 and S_V1F3, and construction for variant c.884A>G replaced S_F4 and S_F5 with S_V2F4 and S_V2F5. For MAPT exon 10, the wildtype was constructed using oligonucleotides M_F1, M_F2, M_F3, M_F4, M_F5, and M_F6. Construction for variant c.1841C>T replaced

M_F3 with M_V1F3, and construction for variant c.del1846_1848 replaced M_F3 with M_V2F3.For DMD exon 29, the wildtype was constructed using oligonucleotides D_F1, D_F2, D_F3, D_F4, D_F5, D_F6, D_F7, and D_F8. Construction for variant c.3940C>T replaced D_F3 with D_V1F3.

The first and last fragment for all exons also contained the universal backbone Gibson overlap sequence, allowing for the assembled fragment to be integrated into a plasmid. These plasmids were transformed into 5-alpha Eletrocompetenet *E. Coli* cells (New England Biolabs), which is a DH5alpha strain modified to reduce recombination of cloned DNA. Once transformed, cells were grown at 37˚C overnight and plasmid was extracted and purified using a miniprep kit (Qiagen).

Each of these eight plasmids were transfected into HEK293 cells using Lipofectamine 3000 Transfection Agent (ThermoFisher) following manufacturer's protocol. Cells were incubated at 37˚C overnight to allow for transcription of plasmids to occur. The following day, cells were lysed and their RNA was purified using an RNeasy mini RNA extraction kit (Qiagen). Purified RNA was then reverse transcribed using Maxima H Minus Reverse Transcriptase (ThermoFisher) following manufacturer's protocol using ESM_006 as reverse transcription primer. This primer binds in the UTR region of the citrine gene just downstream of the citrine exon two. The resulting cDNA was then PCR amplified using ESM_007 and ESM_008. The forward primer (ESM_007) binds in the citrine exon one region, with the reverse primer (ESM_008) binding in the citrine exon two region. Therefore, the PCR amplicon will span the exon-exon junction of the two citrine exons and, if spliced in, will include additional exon-exon junctions with the exon included in the insert region. If the exon included in the insert region is spliced in, the resulting amplicon of this PCR reaction will be much longer. The PCR product is then run on an agarose gel and imaged. In cases

where alternative splicing occurred, two bands will appear, with the top band representing the products where exon of interest was spliced in and the bottom band representing products where the exon of interest was spliced out. Using methods identical to how CleavR enrichment was calculated using gel images (through ImageJ, please refer to this section for more details), it is possible to quantify the PSI and DPSI values for each wildtype and variant of the wildtype tested.

Oligonucleotides Used

SPLiT-seq general oligonucleotides:

Oligonucleotie Number	Description	Sequence
BC_0060	Round 3 barcode linker	AGTCGTACGCCGATGCGAAACATCGGCCAC
BC_0062	PCR primer, used after template switching (used with BC_0108)	CAGACGTGTGCTCTTCCGATCT
BC_0066	Round 3 blocking strand	GTGGCCGATGTTTCGCATCGGCGTACGACT
BC_0076	Nextera Tagmentation PCR primer (TSBC07), sublibrary index #1 (used with BC_0118)	CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCAGACGTGTGC TCTTCCGATCT
BC_0077	Nextera Tagmentation PCR primer (TSBC08), sublibrary index #2 (used with BC_0118)	CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAGTTCAGACGTGTGC TCTTCCGATCT
BC_0078	Nextera Tagmentation PCR primer (TSBC09), sublibrary index #3 (used with BC_0118)	CAAGCAGAAGACGGCATACGAGATCTGATCGTGACTGGAGTTCAGACGTGTGC TCTTCCGATCT
BC_0079	Nextera Tagmentation PCR primer (TSBC10), sublibrary index #4 (used with BC_0118)	CAAGCAGAAGACGGCATACGAGATAAGCTAGTGACTGGAGTTCAGACGTGTGC TCTTCCGATCT
BC_0080	Nextera Tagmentation PCR primer (TSBC11), sublibrary index #5 (used with BC_0118)	CAAGCAGAAGACGGCATACGAGATGTAGCCGTGACTGGAGTTCAGACGTGTGC TCTTCCGATCT
BC_0081	Nextera Tagmentation PCR primer (TSBC12), sublibrary index #6 (used with BC_0118)	CAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAGTTCAGACGTGTGC TCTTCCGATCT
BC_0082	Nextera Tagmentation PCR primer (TSBC13), sublibrary index #7 (used with BC_0118)	CAAGCAGAAGACGGCATACGAGATTTGACTGTGACTGGAGTTCAGACGTGTGCT CTTCCGATCT
BC_0083	Nextera Tagmentation PCR primer (TSBC14), sublibrary index #8 (used with BC_0118)	CAAGCAGAAGACGGCATACGAGATGGAACTGTGACTGGAGTTCAGACGTGTGC TCTTCCGATCT
BC_0108	PCR primer, used after template switching (used with BC_0062)	AAGCAGTGGTATCAACGCAGAGT
BC_0118	Nextera Tagmentation PCR primer N501 (used with BC_0076 through BC_0083)	AATGATACGGCGACCACCGAGATCTACACTAGATCGCTCGTCGGCAGCGTCAGA TGTGTATAAGAGACAG

BC_0127	Template switching primer, HPLC purified (purchased from Exiqon)	AAGCAGTGGTATCAACGCAGAGTGAATrGrG+G
BC_0215	Round 2 barcode linker	CGAATGCTCTGGCCTCTCAAGCACGTGGAT
BC_0216	Round 2 blocking strand	ATCCACGTGCTTGAGAGGCCAGAGCATTCG

SPLiT-seq Round 1 oligonucleotides:

WellPosition	Primer Type	Name	Sequence
A1	dt(15)VN	Round1_01	/5Phos/AGGCCAGAGCATTCGAACGTGATTTTTTTTTTTTTTTVN
A2	dt(15)VN	Round1_02	/5Phos/AGGCCAGAGCATTCGAAACATCGTTTTTTTTTTTTTTTVN
A3	dt(15)VN	Round1_03	/5Phos/AGGCCAGAGCATTCGATGCCTAATTTTTTTTTTTTTTTVN
A4	dt(15)VN	Round1_04	/5Phos/AGGCCAGAGCATTCGAGTGGTCATTTTTTTTTTTTTTTVN
A5	dt(15)VN	Round1_05	/5Phos/AGGCCAGAGCATTCGACCACTGTTTTTTTTTTTTTTTVN
A6	dt(15)VN	Round1_06	/5Phos/AGGCCAGAGCATTCGACATTGGCTTTTTTTTTTTTTTTVN
A7	dt(15)VN	Round1_07	/5Phos/AGGCCAGAGCATTCGCAGATCTGTTTTTTTTTTTTTTTVN
A8	dt(15)VN	Round1_08	/5Phos/AGGCCAGAGCATTCGCATCAAGTTTTTTTTTTTTTTTVN
A9	dt(15)VN	Round1_09	/5Phos/AGGCCAGAGCATTCGCGCTGATCTTTTTTTTTTTTTTTVN
A10	dt(15)VN	Round1_10	/5Phos/AGGCCAGAGCATTCGACAAGCTATTTTTTTTTTTTTTTVN
A11	dt(15)VN	Round1_11	/5Phos/AGGCCAGAGCATTCGCTGTAGCCTTTTTTTTTTTTTTTVN
A12	dt(15)VN	Round1_12	/5Phos/AGGCCAGAGCATTCGAGTACAAGTTTTTTTTTTTTTTTVN
B1	dt(15)VN	Round1_13	/5Phos/AGGCCAGAGCATTCGAACAACCATTTTTTTTTTTTTTTVN
B2	dt(15)VN	Round1_14	/5Phos/AGGCCAGAGCATTCGAACCGAGATTTTTTTTTTTTTTTVN
B3	dt(15)VN	Round1_15	/5Phos/AGGCCAGAGCATTCGAACGCTTATTTTTTTTTTTTTTTVN
B4	dt(15)VN	Round1_16	/5Phos/AGGCCAGAGCATTCGAAGACGGATTTTTTTTTTTTTTTVN
B5	dt(15)VN	Round1_17	/5Phos/AGGCCAGAGCATTCGAAGGTACATTTTTTTTTTTTTTTVN
B6	dt(15)VN	Round1_18	/5Phos/AGGCCAGAGCATTCGACACAGAATTTTTTTTTTTTTTTVN
B7	dt(15)VN	Round1_19	/5Phos/AGGCCAGAGCATTCGACAGCAGATTTTTTTTTTTTTTTVN
B8	dt(15)VN	Round1_20	/5Phos/AGGCCAGAGCATTCGACCTCCAATTTTTTTTTTTTTTTVN
B9	dt(15)VN	Round1_21	/5Phos/AGGCCAGAGCATTCGACGCTCGATTTTTTTTTTTTTTTVN
B10	dt(15)VN	Round1_22	/5Phos/AGGCCAGAGCATTCGACGTATCATTTTTTTTTTTTTTTVN
B11	dt(15)VN	Round1_23	/5Phos/AGGCCAGAGCATTCGACTATGCATTTTTTTTTTTTTTTVN
B12	dt(15)VN	Round1_24	/5Phos/AGGCCAGAGCATTCGAGAGTCAATTTTTTTTTTTTTTTVN
C1	dt(15)VN	Round1_25	/5Phos/AGGCCAGAGCATTCGAGATCGCATTTTTTTTTTTTTTTVN
C2	dt(15)VN	Round1_26	/5Phos/AGGCCAGAGCATTCGAGCAGGAATTTTTTTTTTTTTTTVN
C3	dt(15)VN	Round1_27	/5Phos/AGGCCAGAGCATTCGAGTCACTATTTTTTTTTTTTTTTVN
C4	dt(15)VN	Round1_28	/5Phos/AGGCCAGAGCATTCGATCCTGTATTTTTTTTTTTTTTTVN
C5	dt(15)VN	Round1_29	/5Phos/AGGCCAGAGCATTCGATTGAGGATTTTTTTTTTTTTTTVN
C6	dt(15)VN	Round1_30	/5Phos/AGGCCAGAGCATTCGCAACCACATTTTTTTTTTTTTTTVN
C7	dt(15)VN	Round1_31	/5Phos/AGGCCAGAGCATTCGGACTAGTATTTTTTTTTTTTTTTVN
C8	dt(15)VN	Round1_32	/5Phos/AGGCCAGAGCATTCGCAATGGAATTTTTTTTTTTTTTTVN
C9	dt(15)VN	Round1_33	/5Phos/AGGCCAGAGCATTCGCACTTCGATTTTTTTTTTTTTTTVN
C10	dt(15)VN	Round1_34	/5Phos/AGGCCAGAGCATTCGCAGCGTTATTTTTTTTTTTTTTTVN
C11	dt(15)VN	Round1_35	/5Phos/AGGCCAGAGCATTCGCATACCAATTTTTTTTTTTTTTTVN
C12	dt(15)VN	Round1_36	/5Phos/AGGCCAGAGCATTCGCCAGTTCATTTTTTTTTTTTTTTVN
D1	dt(15)VN	Round1_37	/5Phos/AGGCCAGAGCATTCGCCGAAGTATTTTTTTTTTTTTTTVN
D2	dt(15)VN	Round1_38	/5Phos/AGGCCAGAGCATTCGCCGTGAGATTTTTTTTTTTTTTTVN
D3	dt(15)VN	Round1_39	/5Phos/AGGCCAGAGCATTCGCCTCCTGATTTTTTTTTTTTTTTVN
D4	dt(15)VN	Round1_40	/5Phos/AGGCCAGAGCATTCGCGAACTTATTTTTTTTTTTTTTTVN
D5	dt(15)VN	Round1_41	/5Phos/AGGCCAGAGCATTCGCGACTGGATTTTTTTTTTTTTTTVN
D6	dt(15)VN	Round1_42	/5Phos/AGGCCAGAGCATTCGCGCATACATTTTTTTTTTTTTTTVN
D7	dt(15)VN	Round1_43	/5Phos/AGGCCAGAGCATTCGCTCAATGATTTTTTTTTTTTTTTVN
D8	dt(15)VN	Round1_44	/5Phos/AGGCCAGAGCATTCGCTGAGCCATTTTTTTTTTTTTTTVN
D9	dt(15)VN	Round1_45	/5Phos/AGGCCAGAGCATTCGCTGGCATATTTTTTTTTTTTTTTVN
D10	dt(15)VN	Round1_46	/5Phos/AGGCCAGAGCATTCGGAATCTGATTTTTTTTTTTTTTTVN
D11	dt(15)VN	Round1_47	/5Phos/AGGCCAGAGCATTCGCAAGACTATTTTTTTTTTTTTTTVN
D12	dt(15)VN	Round1_48	/5Phos/AGGCCAGAGCATTCGGAGCTGAATTTTTTTTTTTTTTTVN
E1	random hexamer	Round1_49	/5Phos/AGGCCAGAGCATTCGGATAGACANNNNNN
E2	random hexamer	Round1_50	/5Phos/AGGCCAGAGCATTCGGCCACATANNNNNN
E3	random hexamer	Round1_51	/5Phos/AGGCCAGAGCATTCGGCGAGTAANNNNNN

E4	random hexamer	Round1_52	/5Phos/AGGCCAGAGCATTCGGCTAACGANNNNNN
E5	random hexamer	Round1_53	/5Phos/AGGCCAGAGCATTCGGCTCGGTANNNNNN
E6	random hexamer	Round1_54	/5Phos/AGGCCAGAGCATTCGGGAGAACANNNNNN
E7	random hexamer	Round1_55	/5Phos/AGGCCAGAGCATTCGGGTGCGAANNNNNN
E8	random hexamer	Round1_56	/5Phos/AGGCCAGAGCATTCGGTACGCAANNNNNN
E9	random hexamer	Round1_57	/5Phos/AGGCCAGAGCATTCGGTCGTAGANNNNNN
E10	random hexamer	Round1_58	/5Phos/AGGCCAGAGCATTCGGTCTGTCANNNNNN
E11	random hexamer	Round1_59	/5Phos/AGGCCAGAGCATTCGGTGTTCTANNNNNN
E12	random hexamer	Round1_60	/5Phos/AGGCCAGAGCATTCGTAGGATGANNNNNN
F1	random hexamer	Round1_61	/5Phos/AGGCCAGAGCATTCGTATCAGCANNNNNN
F2	random hexamer	Round1_62	/5Phos/AGGCCAGAGCATTCGTCCGTCTANNNNNN
F3	random hexamer	Round1_63	/5Phos/AGGCCAGAGCATTCGTCTTCACANNNNNN
F4	random hexamer	Round1_64	/5Phos/AGGCCAGAGCATTCGTGAAGAGANNNNNN
F5	random hexamer	Round1_65	/5Phos/AGGCCAGAGCATTCGTGGAACAANNNNNN
F6	random hexamer	Round1_66	/5Phos/AGGCCAGAGCATTCGTGGCTTCANNNNNN
F7	random hexamer	Round1_67	/5Phos/AGGCCAGAGCATTCGTGGTGGTANNNNNN
F8	random hexamer	Round1_68	/5Phos/AGGCCAGAGCATTCGTTCACGCANNNNNN
F9	random hexamer	Round1_69	/5Phos/AGGCCAGAGCATTCGAACTCACCNNNNNN
F10	random hexamer	Round1_70	/5Phos/AGGCCAGAGCATTCGAAGAGATCNNNNNN
F11	random hexamer	Round1_71	/5Phos/AGGCCAGAGCATTCGAAGGACACNNNNNN
F12	random hexamer	Round1_72	/5Phos/AGGCCAGAGCATTCGAATCCGTCNNNNNN
G1	random hexamer	Round1_73	/5Phos/AGGCCAGAGCATTCGAATGTTGCNNNNNN
G2	random hexamer	Round1_74	/5Phos/AGGCCAGAGCATTCGACACGACCNNNNNN
G3	random hexamer	Round1_75	/5Phos/AGGCCAGAGCATTCGACAGATTCNNNNNN
G4	random hexamer	Round1_76	/5Phos/AGGCCAGAGCATTCGAGATGTACNNNNNN
G5	random hexamer	Round1_77	/5Phos/AGGCCAGAGCATTCGAGCACCTCNNNNNN
G6	random hexamer	Round1_78	/5Phos/AGGCCAGAGCATTCGAGCCATGCNNNNNN
G7	random hexamer	Round1_79	/5Phos/AGGCCAGAGCATTCGAGGCTAACNNNNNN
G8	random hexamer	Round1_80	/5Phos/AGGCCAGAGCATTCGATAGCGACNNNNNN
G9	random hexamer	Round1_81	/5Phos/AGGCCAGAGCATTCGATCATTCCNNNNNN
G10	random hexamer	Round1_82	/5Phos/AGGCCAGAGCATTCGATTGGCTCNNNNNN
G11	random hexamer	Round1_83	/5Phos/AGGCCAGAGCATTCGCAAGGAGCNNNNNN
G12	random hexamer	Round1_84	/5Phos/AGGCCAGAGCATTCGCACCTTACNNNNNN
H1	random hexamer	Round1_85	/5Phos/AGGCCAGAGCATTCGCCATCCTCNNNNNN
H2	random hexamer	Round1_86	/5Phos/AGGCCAGAGCATTCGCCGACAACNNNNNN
H3	random hexamer	Round1_87	/5Phos/AGGCCAGAGCATTCGCCTAATCCNNNNNN
H4	random hexamer	Round1_88	/5Phos/AGGCCAGAGCATTCGCCTCTATCNNNNNN
H5	random hexamer	Round1_89	/5Phos/AGGCCAGAGCATTCGCGACACACNNNNNN
H6	random hexamer	Round1_90	/5Phos/AGGCCAGAGCATTCGCGGATTGCNNNNNN
H7	random hexamer	Round1_91	/5Phos/AGGCCAGAGCATTCGCTAAGGTCNNNNNN
H8	random hexamer	Round1_92	/5Phos/AGGCCAGAGCATTCGGAACAGGCNNNNNN
H9	random hexamer	Round1_93	/5Phos/AGGCCAGAGCATTCGGACAGTGCNNNNNN
H10	random hexamer	Round1_94	/5Phos/AGGCCAGAGCATTCGGAGTTAGCNNNNNN
H11	random hexamer	Round1_95	/5Phos/AGGCCAGAGCATTCGGATGAATCNNNNNN
H12	random hexamer	Round1_96	/5Phos/AGGCCAGAGCATTCGGCCAAGACNNNNNN

SPLiT-seq Round 2 oligonucleotides

WellPosition	Name	Sequence
A1	Round2_01	/5Phos/CATCGGCGTACGACTAACGTGATATCCACGTGCTTGAG
A2	Round2_02	/5Phos/CATCGGCGTACGACTAAACATCGATCCACGTGCTTGAG
A3	Round2_03	/5Phos/CATCGGCGTACGACTATGCCTAAATCCACGTGCTTGAG
A4	Round2_04	/5Phos/CATCGGCGTACGACTAGTGGTCAATCCACGTGCTTGAG
A5	Round2_05	/5Phos/CATCGGCGTACGACTACCACTGTATCCACGTGCTTGAG
A6	Round2_06	/5Phos/CATCGGCGTACGACTACATTGGCATCCACGTGCTTGAG
A7	Round2_07	/5Phos/CATCGGCGTACGACTCAGATCTGATCCACGTGCTTGAG
A8	Round2_08	/5Phos/CATCGGCGTACGACTCATCAAGTATCCACGTGCTTGAG
A9	Round2_09	/5Phos/CATCGGCGTACGACTCGCTGATCATCCACGTGCTTGAG

A10	Round2_10	/5Phos/CATCGGCGTACGACTACAAGCTAATCCACGTGCTTGAG
A11	Round2_11	/5Phos/CATCGGCGTACGACTCTGTAGCCATCCACGTGCTTGAG
A12	Round2_12	/5Phos/CATCGGCGTACGACTAGTACAAGATCCACGTGCTTGAG
B1	Round2_13	/5Phos/CATCGGCGTACGACTAACAACCAATCCACGTGCTTGAG
B2	Round2_14	/5Phos/CATCGGCGTACGACTAACCGAGAATCCACGTGCTTGAG
B3	Round2_15	/5Phos/CATCGGCGTACGACTAACGCTTAATCCACGTGCTTGAG
B4	Round2_16	/5Phos/CATCGGCGTACGACTAAGACGGAATCCACGTGCTTGAG
B5	Round2_17	/5Phos/CATCGGCGTACGACTAAGGTACAATCCACGTGCTTGAG
B6	Round2_18	/5Phos/CATCGGCGTACGACTACACAGAAATCCACGTGCTTGAG
B7	Round2_19	/5Phos/CATCGGCGTACGACTACAGCAGAATCCACGTGCTTGAG
B8	Round2_20	/5Phos/CATCGGCGTACGACTACCTCCAAATCCACGTGCTTGAG
B9	Round2_21	/5Phos/CATCGGCGTACGACTACGCTCGAATCCACGTGCTTGAG
B10	Round2_22	/5Phos/CATCGGCGTACGACTACGTATCAATCCACGTGCTTGAG
B11	Round2_23	/5Phos/CATCGGCGTACGACTACTATGCAATCCACGTGCTTGAG
B12	Round2_24	/5Phos/CATCGGCGTACGACTAGAGTCAAATCCACGTGCTTGAG
C1	Round2_25	/5Phos/CATCGGCGTACGACTAGATCGCAATCCACGTGCTTGAG
C2	Round2_26	/5Phos/CATCGGCGTACGACTAGCAGGAAATCCACGTGCTTGAG
C3	Round2_27	/5Phos/CATCGGCGTACGACTAGTCACTAATCCACGTGCTTGAG
C4	Round2_28	/5Phos/CATCGGCGTACGACTATCCTGTAATCCACGTGCTTGAG
C5	Round2_29	/5Phos/CATCGGCGTACGACTATTGAGGAATCCACGTGCTTGAG
C6	Round2_30	/5Phos/CATCGGCGTACGACTCAACCACAATCCACGTGCTTGAG
C7	Round2_31	/5Phos/CATCGGCGTACGACTGACTAGTAATCCACGTGCTTGAG
C8	Round2_32	/5Phos/CATCGGCGTACGACTCAATGGAAATCCACGTGCTTGAG
C9	Round2_33	/5Phos/CATCGGCGTACGACTCACTTCGAATCCACGTGCTTGAG
C10	Round2_34	/5Phos/CATCGGCGTACGACTCAGCGTTAATCCACGTGCTTGAG
C11	Round2_35	/5Phos/CATCGGCGTACGACTCATACCAAATCCACGTGCTTGAG
C12	Round2_36	/5Phos/CATCGGCGTACGACTCCAGTTCAATCCACGTGCTTGAG
D1	Round2_37	/5Phos/CATCGGCGTACGACTCCGAAGTAATCCACGTGCTTGAG
D2	Round2_38	/5Phos/CATCGGCGTACGACTCCGTGAGAATCCACGTGCTTGAG
D3	Round2_39	/5Phos/CATCGGCGTACGACTCCTCCTGAATCCACGTGCTTGAG
D4	Round2_40	/5Phos/CATCGGCGTACGACTCGAACTTAATCCACGTGCTTGAG
D5	Round2_41	/5Phos/CATCGGCGTACGACTCGACTGGAATCCACGTGCTTGAG
D6	Round2_42	/5Phos/CATCGGCGTACGACTCGCATACAATCCACGTGCTTGAG
D7	Round2_43	/5Phos/CATCGGCGTACGACTCTCAATGAATCCACGTGCTTGAG
D8	Round2_44	/5Phos/CATCGGCGTACGACTCTGAGCCAATCCACGTGCTTGAG
D9	Round2_45	/5Phos/CATCGGCGTACGACTCTGGCATAATCCACGTGCTTGAG
D10	Round2_46	/5Phos/CATCGGCGTACGACTGAATCTGAATCCACGTGCTTGAG
D11	Round2_47	/5Phos/CATCGGCGTACGACTCAAGACTAATCCACGTGCTTGAG
D12	Round2_48	/5Phos/CATCGGCGTACGACTGAGCTGAAATCCACGTGCTTGAG
E1	Round2_49	/5Phos/CATCGGCGTACGACTGATAGACAATCCACGTGCTTGAG
E2	Round2_50	/5Phos/CATCGGCGTACGACTGCCACATAATCCACGTGCTTGAG
E3	Round2_51	/5Phos/CATCGGCGTACGACTGCGAGTAAATCCACGTGCTTGAG
E4	Round2_52	/5Phos/CATCGGCGTACGACTGCTAACGAATCCACGTGCTTGAG
E5	Round2_53	/5Phos/CATCGGCGTACGACTGCTCGGTAATCCACGTGCTTGAG
E6	Round2_54	/5Phos/CATCGGCGTACGACTGGAGAACAATCCACGTGCTTGAG
E7	Round2_55	/5Phos/CATCGGCGTACGACTGGTGCGAAATCCACGTGCTTGAG
E8	Round2_56	/5Phos/CATCGGCGTACGACTGTACGCAAATCCACGTGCTTGAG
E9	Round2_57	/5Phos/CATCGGCGTACGACTGTCGTAGAATCCACGTGCTTGAG
E10	Round2_58	/5Phos/CATCGGCGTACGACTGTCTGTCAATCCACGTGCTTGAG
E11	Round2_59	/5Phos/CATCGGCGTACGACTGTGTTCTAATCCACGTGCTTGAG
E12	Round2_60	/5Phos/CATCGGCGTACGACTTAGGATGAATCCACGTGCTTGAG
F1	Round2_61	/5Phos/CATCGGCGTACGACTTATCAGCAATCCACGTGCTTGAG
F2	Round2_62	/5Phos/CATCGGCGTACGACTTCCGTCTAATCCACGTGCTTGAG
F3	Round2_63	/5Phos/CATCGGCGTACGACTTCTTCACAATCCACGTGCTTGAG
F4	Round2_64	/5Phos/CATCGGCGTACGACTTGAAGAGAATCCACGTGCTTGAG
F5	Round2_65	/5Phos/CATCGGCGTACGACTTGGAACAAATCCACGTGCTTGAG
F6	Round2_66	/5Phos/CATCGGCGTACGACTTGGCTTCAATCCACGTGCTTGAG
F7	Round2_67	/5Phos/CATCGGCGTACGACTTGGTGGTAATCCACGTGCTTGAG
F8	Round2_68	/5Phos/CATCGGCGTACGACTTTCACGCAATCCACGTGCTTGAG
F9	Round2_69	/5Phos/CATCGGCGTACGACTAACTCACCATCCACGTGCTTGAG
F10	Round2_70	/5Phos/CATCGGCGTACGACTAAGAGATCATCCACGTGCTTGAG
F11	Round2_71	/5Phos/CATCGGCGTACGACTAAGGACACATCCACGTGCTTGAG
F12	Round2_72	/5Phos/CATCGGCGTACGACTAATCCGTCATCCACGTGCTTGAG
G1	Round2_73	/5Phos/CATCGGCGTACGACTAATGTTGCATCCACGTGCTTGAG

G2	Round2_74	/5Phos/CATCGGCGTACGACTACACGACCATCCACGTGCTTGAG
G3	Round2_75	/5Phos/CATCGGCGTACGACTACAGATTCATCCACGTGCTTGAG
G4	Round2_76	/5Phos/CATCGGCGTACGACTAGATGTACATCCACGTGCTTGAG
G5	Round2_77	/5Phos/CATCGGCGTACGACTAGCACCTCATCCACGTGCTTGAG
G6	Round2_78	/5Phos/CATCGGCGTACGACTAGCCATGCATCCACGTGCTTGAG
G7	Round2_79	/5Phos/CATCGGCGTACGACTAGGCTAACATCCACGTGCTTGAG
G8	Round2_80	/5Phos/CATCGGCGTACGACTATAGCGACATCCACGTGCTTGAG
G9	Round2_81	/5Phos/CATCGGCGTACGACTATCATTCCATCCACGTGCTTGAG
G10	Round2_82	/5Phos/CATCGGCGTACGACTATTGGCTCATCCACGTGCTTGAG
G11	Round2_83	/5Phos/CATCGGCGTACGACTCAAGGAGCATCCACGTGCTTGAG
G12	Round2_84	/5Phos/CATCGGCGTACGACTCACCTTACATCCACGTGCTTGAG
H1	Round2_85	/5Phos/CATCGGCGTACGACTCCATCCTCATCCACGTGCTTGAG
H2	Round2_86	/5Phos/CATCGGCGTACGACTCCGACAACATCCACGTGCTTGAG
H3	Round2_87	/5Phos/CATCGGCGTACGACTCCTAATCCATCCACGTGCTTGAG
H4	Round2_88	/5Phos/CATCGGCGTACGACTCCTCTATCATCCACGTGCTTGAG
H5	Round2_89	/5Phos/CATCGGCGTACGACTCGACACACATCCACGTGCTTGAG
H6	Round2_90	/5Phos/CATCGGCGTACGACTCGGATTGCATCCACGTGCTTGAG
H7	Round2_91	/5Phos/CATCGGCGTACGACTCTAAGGTCATCCACGTGCTTGAG
H8	Round2_92	/5Phos/CATCGGCGTACGACTGAACAGGCATCCACGTGCTTGAG
H9	Round2_93	/5Phos/CATCGGCGTACGACTGACAGTGCATCCACGTGCTTGAG
H10	Round2_94	/5Phos/CATCGGCGTACGACTGAGTTAGCATCCACGTGCTTGAG
H11	Round2_95	/5Phos/CATCGGCGTACGACTGATGAATCATCCACGTGCTTGAG
H12	Round2_96	/5Phos/CATCGGCGTACGACTGCCAAGACATCCACGTGCTTGAG

SPLiT-seq Round 3 oligonucleotides:

WellPosition	Name	Sequence
A1	Round3_01	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAACGTGATGTGGCCGATGTTTCG
A2	Round3_02	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAAACATCGGTGGCCGATGTTTCG
A3	Round3_03	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNATGCCTAAGTGGCCGATGTTTCG
A4	Round3_04	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAGTGGTCAGTGGCCGATGTTTCG
A5	Round3_05	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNACCACTGTGTGGCCGATGTTTCG
A6	Round3_06	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNACATTGGCGTGGCCGATGTTTCG
A7	Round3_07	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCAGATCTGGTGGCCGATGTTTCG
A8	Round3_08	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCATCAAGTGTGGCCGATGTTTCG
A9	Round3_09	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCGCTGATCGTGGCCGATGTTTCG
A10	Round3_10	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNACAAGCTAGTGGCCGATGTTTCG
A11	Round3_11	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCTGTAGCCGTGGCCGATGTTTCG
A12	Round3_12	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAGTACAAGGTGGCCGATGTTTCG
B1	Round3_13	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAACAACCAGTGGCCGATGTTTCG
B2	Round3_14	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAACCGAGAGTGGCCGATGTTTCG
B3	Round3_15	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAACGCTTAGTGGCCGATGTTTCG
B4	Round3_16	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAAGACGGAGTGGCCGATGTTTCG
B5	Round3_17	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAAGGTACAGTGGCCGATGTTTCG
B6	Round3_18	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNACACAGAAGTGGCCGATGTTTCG
B7	Round3_19	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNACAGCAGAGTGGCCGATGTTTCG
B8	Round3_20	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNACCTCCAAGTGGCCGATGTTTCG
B9	Round3_21	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNACGCTCGAGTGGCCGATGTTTCG
B10	Round3_22	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNACGTATCAGTGGCCGATGTTTCG
B11	Round3_23	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNACTATGCAGTGGCCGATGTTTCG
B12	Round3_24	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAGAGTCAAGTGGCCGATGTTTCG
C1	Round3_25	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAGATCGCAGTGGCCGATGTTTCG
C2	Round3_26	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAGCAGGAAGTGGCCGATGTTTCG
C3	Round3_27	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAGTCACTAGTGGCCGATGTTTCG
C4	Round3_28	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNATCCTGTAGTGGCCGATGTTTCG
C5	Round3_29	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNATTGAGGAGTGGCCGATGTTTCG
C6	Round3_30	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCAACCACAGTGGCCGATGTTTCG
C7	Round3_31	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGACTAGTAGTGGCCGATGTTTCG
C8	Round3_32	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCAATGGAAGTGGCCGATGTTTCG
C9	Round3_33	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCACTTCGAGTGGCCGATGTTTCG

C10	Round3_34	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCAGCGTTAGTGGCCGATGTTTCG
C11	Round3_35	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCATACCAAGTGGCCGATGTTTCG
C12	Round3_36	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCCAGTTCAGTGGCCGATGTTTCG
D1	Round3_37	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCCGAAGTAGTGGCCGATGTTTCG
D2	Round3_38	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCCGTGAGAGTGGCCGATGTTTCG
D3	Round3_39	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCCTCCTGAGTGGCCGATGTTTCG
D4	Round3_40	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCGAACTTAGTGGCCGATGTTTCG
D5	Round3_41	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCGACTGGAGTGGCCGATGTTTCG
D6	Round3_42	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCGCATACAGTGGCCGATGTTTCG
D7	Round3_43	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCTCAATGAGTGGCCGATGTTTCG
D8	Round3_44	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCTGAGCCAGTGGCCGATGTTTCG
D9	Round3_45	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCTGGCATAGTGGCCGATGTTTCG
D10	Round3_46	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGAATCTGAGTGGCCGATGTTTCG
D11	Round3_47	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCAAGACTAGTGGCCGATGTTTCG
D12	Round3_48	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGAGCTGAAGTGGCCGATGTTTCG
E1	Round3_49	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGATAGACAGTGGCCGATGTTTCG
E2	Round3_50	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGCCACATAGTGGCCGATGTTTCG
E3	Round3_51	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGCGAGTAAGTGGCCGATGTTTCG
E4	Round3_52	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGCTAACGAGTGGCCGATGTTTCG
E5	Round3_53	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGCTCGGTAGTGGCCGATGTTTCG
E6	Round3_54	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGGAGAACAGTGGCCGATGTTTCG
E7	Round3_55	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGGTGCGAAGTGGCCGATGTTTCG
E8	Round3_56	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGTACGCAAGTGGCCGATGTTTCG
E9	Round3_57	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGTCGTAGAGTGGCCGATGTTTCG
E10	Round3_58	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGTCTGTCAGTGGCCGATGTTTCG
E11	Round3_59	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGTGTTCTAGTGGCCGATGTTTCG
E12	Round3_60	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNTAGGATGAGTGGCCGATGTTTCG
F1	Round3_61	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNTATCAGCAGTGGCCGATGTTTCG
F2	Round3_62	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNTCCGTCTAGTGGCCGATGTTTCG
F3	Round3_63	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNTCTTCACAGTGGCCGATGTTTCG
F4	Round3_64	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNTGAAGAGAGTGGCCGATGTTTCG
F5	Round3_65	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNTGGAACAAGTGGCCGATGTTTCG
F6	Round3_66	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNTGGCTTCAGTGGCCGATGTTTCG
F7	Round3_67	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNTGGTGGTAGTGGCCGATGTTTCG
F8	Round3_68	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNTTCACGCAGTGGCCGATGTTTCG
F9	Round3_69	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAACTCACCGTGGCCGATGTTTCG
F10	Round3_70	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAAGAGATCGTGGCCGATGTTTCG
F11	Round3_71	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAAGGACACGTGGCCGATGTTTCG
F12	Round3_72	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAATCCGTCGTGGCCGATGTTTCG
G1	Round3_73	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAATGTTGCGTGGCCGATGTTTCG
G2	Round3_74	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNACACGACCGTGGCCGATGTTTCG
G3	Round3_75	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNACAGATTCGTGGCCGATGTTTCG
G4	Round3_76	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAGATGTACGTGGCCGATGTTTCG
G5	Round3_77	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAGCACCTCGTGGCCGATGTTTCG
G6	Round3_78	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAGCCATGCGTGGCCGATGTTTCG
G7	Round3_79	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAGGCTAACGTGGCCGATGTTTCG
G8	Round3_80	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNATAGCGACGTGGCCGATGTTTCG
G9	Round3_81	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNATCATTCCGTGGCCGATGTTTCG
G10	Round3_82	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNATTGGCTCGTGGCCGATGTTTCG
G11	Round3_83	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCAAGGAGCGTGGCCGATGTTTCG
G12	Round3_84	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCACCTTACGTGGCCGATGTTTCG
H1	Round3_85	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCCATCCTCGTGGCCGATGTTTCG
H2	Round3_86	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCCGACAACGTGGCCGATGTTTCG
H3	Round3_87	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCCTAATCCGTGGCCGATGTTTCG
H4	Round3_88	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCCTCTATCGTGGCCGATGTTTCG
H5	Round3_89	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCGACACACGTGGCCGATGTTTCG
H6	Round3_90	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCGGATTGCGTGGCCGATGTTTCG
H7	Round3_91	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCTAAGGTCGTGGCCGATGTTTCG
H8	Round3_92	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGAACAGGCGTGGCCGATGTTTCG
H9	Round3_93	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGACAGTGCGTGGCCGATGTTTCG
H10	Round3_94	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGAGTTAGCGTGGCCGATGTTTCG
H11	Round3_95	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGATGAATCGTGGCCGATGTTTCG
H12	Round3_96	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGCCAAGACGTGGCCGATGTTTCG

CleavR oligonucleotides

Oligo Number	Description	Sequence
BC_0301	Hygro strand caputre primer	AGAAGTACTCGCCGATAGTGGAAACCGA
BC_0306	Amp strand capture primer	ACGGGGAGTCAGGCAACTATGGATGA
BC_0357	Fwd PCR primer to generate Hygro strand	TTTTTTTTTTAAGCAGTGGTATCAACrGCArGAGTrGAATGGGTACCAAACGACGAGCGTGACA
BC_0358	Rev PCR primer to generate Hygro strand	TTTTTTTTTTCAGACrGTGTrGCTCTTCCrGATCTATTCCTTTGCCCTCGGACG
BC_0359	Fwd PCR primer to generate Amp strand	TTTTTTTTTTAAGCAGTGGTATCAACrGCArGAGTrGAATGGGTACCAAACGACGAGCGTGACA
BC_0360	Rev PCR primer to generate Amp strand	TTTTTTTTTTCAGACrGTGTrGCTCTTCCrGATCTCCAATGCTTAATCAGTGAGGCACC
BC_0385	Fwd PCR primer for SPLiT-seq libraries	TTTTTTTTTTCAGACrGTrGTrGCTCTTCCrGATCT
BC_0386	Rev PCR primer for SPLiT-seq libraries	TTTTTTTTTTAAGCAGTGrGTATCAACrGCArGArGT
BC_0391	TRAC capture primer	AGAACCCTGACCCTGCCG
BC_0392	TRBC1 capture primer	CTGAAAAACGTGTTCCCACCCGAG
BC_0393	TRBC2 capture primer	ACCTGAACAAGGTGTTCCCACC

Splicing MPRA oligonucleotides

Oligo Number	Description	Sequence
ESM_007	Forward primer for PCR	CAGAAGAACGGCATCAAAGTGA
ESM_008	Reverse primer for PCR	GTCCGCCCTGAGCAAAGAC
ESM_009	Reverse transcription primer	CAGATGGCTGGCAACTAG
ESM_S_F1	SMN2 exon 7 WT Fragment 1	AGGAAGTTAAAAAAAATAGCTATATAGATATAGATAGCTATATATAGATAGCTTTATATGGA
ESM_ S_F2	SMN2 exon 7 WT Fragment 2	TATAGCTATTTTTTTTAACTTCCTTTATTTTCCTTACAGGGTTTTAGAC
ESM_S_F3	SMN2 exon 7 WT Fragment 3	ACCTTCCTTCTTTTTGATTTTGTCTAAAACCCTGTAAGGAAAATAA
ESM_S_F4	SMN2 exon 7 WT Fragment 4	AAAATCAAAAAGAAGGAAGGTGCTCACATTCCTTAAATTAAGG
ESM_S_F5	SMN2 exon 7 WT Fragment 5	CATAATGCTGGCAGACTTACTCCTTAATTTAAGGAATGTGAGC
ESM_S_F6	SMN2 exon 7 WT Fragment 6	AGTAAGTCTGCCAGCATTATGAAAGTGAATCTTACTTTTGTAAAACTTTATG

ESM_S_V1F2	SMN2 exon 7 Fragment 2 for variant 1	ATCTATATCTATATAGCTATTTTTTTTAACTTCCTTTATTTTCCTTACAGGGTTTCAGAC
ESM_S_V1F3	SMN2 exon 7 Fragment 3 for variant 1	ACCTTCCTTCTTTTTGATTTTGTCTGAAACCCTGTAAGGAAAATAA
ESM_S_V2F4	SMN2 exon 7 Fragment 4 for variant 2	AAAATCAAAAAGAAGGAAGGTGCTCACATTCCTTAAATTAGGG
ESM_S_V2F5	SMN2 exon 7 Fragment 5 for variant 2	CATAATGCTGGCAGACTTACTCCCTAATTTAAGGAATGTGAGC
ESM_M_F1	MAPT exon 10 WT Fragment 1	CCTGGACCCGCCTTAGATATAGATAGCTATATATAGATAGCTTTATATGGA
ESM_M_F2	MAPT exon 10 WT Fragment 2	AGGCGGGTCCAGGGTGGCGTGTCACTCATCCTTTTTTCTGGCTACCAAAGGTGCAGATAA
ESM_M_F3	MAPT exon 10 WT Fragment 3	CACTTGGACTGGACGTTGCTAAGATCCAGCTTCTTATTAATTATCTGCACCTTTGGTAGC
ESM_M_F4	MAPT exon 10 WT Fragment 4	CAACGTCCAGTCCAAGTGTGGCTCAAAGGATAATATCAAACACGTC
ESM_M_F5	MAPT exon 10 WT Fragment 5	GGACGTGTGAAGGTACTCACACTGCCGCCTCCCGGGACGTGTTTGATATTATCCTTTG
ESM_M_F6	MAPT exon 10 WT Fragment 6	GTGAGTACCTTCACACGTCCAAAGTGAATCTTACTTTTGTAAAACTTTATG
ESM_M_V1F3	MAPT exon 10 Fragment 3 for variant 1	CACTTGGACTGGACGTTGCTAAGATCCAGCTTCTTCTTAATTATCTGCACCTTTGGTAGC
ESM_M_V2F3	MAPT exon 10 Fragment 3 for variant 2	CACTTGGACTGGACGTTGCTAAGATCCAGCTTATTAATTATCTGCACCTTTGGTAGC
ESM_D_F1	DMD exon 29 WT Fragment 1	TCTCCTTTTTTTTCTAAATACATAGATATAGATAGCTATATATAGATAGCTTTATATGGA
ESM_D_F2	DMD exon 29 WT Fragment 2	TGTATTTAGAAAAAAAAGGAGAAATAGTAATTATTGCAAATGTGTTTCAGTCACTTGAAA
ESM_D_F3	DMD exon 29 WT Fragment 3	AATCTGATTTGGGTTATCCTCTGAATGTCGCATCAAATTTTCAAGTGACTGAAACACATT
ESM_D_F4	DMD exon 29 WT Fragment 4	AGAGGATAACCCAAATCAGATTCGCATATTGGCACAGACCCTAACAGATGGCGGAGTCAT
ESM_D_F5	DMD exon 29 WT Fragment 5	AAATGTCTCAAGTTCCTCATTGATTAGCTCATCCATGACTCCGCCATCTGT
ESM_D_F6	DMD exon 29 WT Fragment 6	CAATGAGGAACTTGAGACATTTAATTCTCGTTGGAGGGAACTACAT
ESM_D_F7	DMD exon 29 WT Fragment 7	TTTTTCACTTATCTTCATACCTCTTCATGTAGTTCCCTCCAACGAGAATT
ESM_D_F8	DMD exon 29 WT Fragment 8	GAAGAGGTATGAAGATAAGTGAAAAAAAAGTGAATCTTACTTTTGTAAAACTTTATG
ESM_D_V1F3	DMD exon 29 Fragment 3 for variant 1	AATCTGATTTGGGTTATCCTCTGAATGTCACATCAAATTTTCAAGTGACTGAAACACATT

Computational Methods

Alignment and generation of cell-gene matrices

To simplify analysis, we first removed any dephased reads in our library (last 6 bases of read did not match the expected sequence). Reads were then filtered based on quality score in the UMI region. Any read with >1 low-quality base (phred <=10) were discarded. Reads with more than one mismatch in any of the three 8 nt cell barcodes were also discarded. The cDNA reads (Read 1) were then mapped to either a combined mm10-hg19 genome or the mm10 genome using STAR (*100*). The aligned reads in the resulting bam file were mapped to exons and genes using TagReadWithGeneExon from the drop-seq tools (*6*). We only considered the primary alignments. Reads that mapped to a gene, but no exon, were considered intronic. Reads mapping to no gene were considered intergenic. We then used Starcode (*101*) to collapse UMIs of aligned reads that were within 1 nt mismatch of another UMI, assuming the two aligned reads were also from the same UBC. Each original barcoded cDNA molecule is amplified before tagmentation and subsequent PCR, so a single UMI-UBC combination can have several distinct cDNA reads corresponding to different parts of the transcript. Occasionally STAR will map these different reads to different genes. As a result, we chose the most frequently assigned gene as the mapping for the given UMI-UBC combination. We then generated a matrix of gene counts for each cell (N x K matrix, with N cells and K genes). For each gene, both intronic and exonic UMI counts were used.

Selecting high quality transcriptomes from the mouse CNS experiment

We discarded any transcriptomes with >1% reads mapping to mt-RNA, to ensure that all of our transcriptomes originated from nuclei. Transcriptomes with fewer than 250 expressed genes or greater than 2,500 expressed genes were also discarded. This resulted in retention of 163,069

transcriptomes. After clustering (see below), cells in putative doublet clusters were filtered as well, yielding 156,049 transcriptomes used for downstream analysis.

Hierarchical clustering of nuclei from the mouse CNS

Cells that passed the QC were clustered using an iterative clustering pipeline described in previous studies (*10, 102*), with adaption for sparse datasets with large numbers of cells. Briefly, cells were clustered in an iterative top down approach. Each clustering iteration consists of three key steps: high variance gene selection, dimensionality reduction, and clustering. To choose high variance genes, we first fitted a loess regression curve between average scaled gene counts and dispersion (variance divided by mean). The regression residuals are then fitted by a normal distribution based on 25% and 75% quantiles to calculate p-values and adjusted p-value. High variance genes with adjusted p-values smaller than 0.1 were used to compute principle components. The proportion of variance for all PCs were converted to Z-values, and PCs with Z-values greater than two were selected for clustering. The Jaccard-Louvain algorithm (*103*) was then used for clustering, which first computes the k-nearest-neighbors (k=15) for each cell based on reduced PCs, then constructs the cell-cell similarity matrix with the Jaccard index based on the number of shared neighbors between every cell, and finally performs clustering using the Louvain algorithm. To make sure the resulting clusters all had distinguishable transcriptomic signatures, we calculated the differentially expressed genes (DEG) for every pair of clusters. A pair of clusters are separable if the deScore, defined as sum of -log10 (adjusted p-value) for all DEG (fold change >2 and adjusted p-value < 0.01, present in >40% cells in foreground, and foreground vs background enrichment ratio is greater than 3.3) is greater than 150 (every gene can contribute maximal score of 20). We merged the nearest pair of clusters that did not pass the above criterion iteratively until all clusters were separable.

We then iteratively applied the same steps described above to each cluster identified from the first clustering iteration. This iterative process was repeated until no further partitions were found. It was possible that clusters derived from different parent clusters could be similar to each other. Therefore, we computed pairwise DEG again, but reduced the threshold deScore threshold to 80 to prevent over agglomeration. This resulted in 98 clusters.

Clusters consisting of less than 40 cells were discarded (5 clusters consisting of 141 total cells). Putative doublet clusters (clusters in which many transcriptomes were generated from doublets between two different cell types) were identified by searching for co-expression of known markers of different cell types (*e.g.* the neuronal marker *Meg3* (*104*) and the oligodendrocyte marker *Mbp* (*12*)). This resulted in the identification of 12 clusters that were likely generated from doublet transcriptomes. These 12 clusters, consisting of 6,878 cells, were then discarded from further analysis.

Finally, we applied a more stringent test of differential expression between clusters. Using previously described criteria (*6*), we merged pairs of clusters with less than 10 differentially expressed genes (>1 natural log difference between clusters and expressed in >20% of cells in one of the two clusters). This procedure resulted in 8 clusters merging into other clusters, yielding 73 final clusters (156,049 nuclei).

PCA and t-distributed Stochastic Neighbor Embedding

We first normalized the matrix of UMI counts. For each cell, we divided the UMI counts by the total number of UMIs per cell. We subtracted the mean from each gene and then divided by the standard deviation of each gene.

We selected a subset of genes on which to perform PCA (for the mouse CNS analysis we selected genes with at least 4 UMI counts in 10 or more transcriptomes). PCA was performed on the normalized matrix using TruncatedSVD from the python package scikit-learn (*105*). For the mouse CNS analysis, we retained the first 100 PCA components and then performed t-distributed Stochastic Neighbor Embedding using a Matlab implementation of the barnes-hut t-SNE algorithm (*106*).

Lineage analysis

While previous work has identified developmental trajectories using scRNA-seq, this has mostly been confined to *in vitro* differentiation experiments. Finding differentiation trajectories in our dataset first required grouping clusters together that might form a putative lineage. To do this we followed the following procedure:

1. *Find clusters near one another in the original t-SNE embedding (Fig. 2A) that seem to form elongated structures.* In a developmental lineage, we expect cells at the start of the lineage to have substantially different transcriptomes than those at the end of the lineage, leading to this "stretching" of the lineage in the t-SNE, with intermediate cells connecting these early and late cells.
2. *Confirm that transcriptomes from the P2 mouse and P11 mouse segregate towards opposite ends of the putative lineage in the t-SNE embedding (Fig. S6).* This gives us additional confidence that the primary variance in gene expression across the putative lineage does in fact correspond to developmental stages rather than some other factor (*e.g.* regional origins).
3. *Re-embed the clusters in the putative lineage with PCA and t-SNE.* For each putative lineage, we redid PCA and t-SNE with just transcriptomes in the clusters forming the

putative lineage. We did this to ensure the ordering of transcriptomes was driven only by expression of relevant genes (*e.g.* so neuronal PCA components do not drive oligodendrocyte ordering). This analysis resulted in the vast majority of re-embeded transcriptomes forming one connected group, with a small number of cells forming other distinct, new clusters. We used the density-based DBSCAN algorithm (*107*) to identify this main lineage and to discard the transcriptomes from these smaller clusters.

4. *Measure gene expression along pseudotime curve.* Each of the lineages we analyzed in this work again formed elongated connected structures when we re-embedded transcriptomes using t-SNE. We then determined the order of transcriptomes through the t-SNE, by projecting them onto a manually drawn curve spanning the entire embedding. The moving average of gene expression was then calculated across these ordered transcriptomes.
5. *Confirm the identity of lineages using known gene markers from literature.* For each lineage, we further validated it using previously characterized marker genes from literature in addition to *in situ* hybridization data from the Allen Brain Institute (*31*).

Generating composite ISH maps for each cluster

We downloaded *in situ* hybridization (ISH) data collected by the Allen Brain Institute from the Kharchenko lab for P4 and P14 mice (*31*). The data consists of 2,187 gene measurements compiled into a 3D (P4: 50 x 43 x 77, P14: 50 x 40 x 68) map of expression "energies" corresponding to ISH staining intensities in each voxel. We used the Allen brain structure annotations to mask any voxels outside annotated brain structures. For voxels with missing data for a given gene, we set the voxel energies to the mean energy for that gene.

We then used the Allen ISH data to create a composite map of differentially expressed genes for each cluster. For each cluster, the top 5 enriched genes (with ISH data) were determined using differential gene expression described above. We normalized the expression of each gene by dividing the intensity in each voxel by the average intensity across all the voxels. To generate a composite map, we then averaged the intensity across all 5 genes within each voxel. To visualize the 3D map in a single image, we summed across sagittal slice numbers 7-24 (out of 50 slices) of the 3D map. Only slices 7-24 were used because many genes did not contain data in the other slices. Genes used to generate these maps can be found in Table S6 and S7.

Re-clustering spinal cord transcriptomes

In our original clustering, over 60% of transcriptomes from the spinal cord clustered into a large unresolved cluster (Fig. S19). Given that we had ~6x more brain nuclei relative to spinal cord nuclei, it is not surprising that the majority of PCA components selected for clustering describe variance in gene expression in the brain rather than the spine. PCA components explaining expression differences in the spinal cord may have been filtered out as "not significant" (Z-value<2).

Therefore, we reasoned that we might be able to distinguish more cell types in the spinal cord if we re-clustered only the spinal cord transcriptomes. For this clustering, we chose to use the Monocle 2 package (*108*). As suggested in the Monocle manual, we first selected high-variance genes (with high dispersion), before performing PCA. We then used the first 50 components as input to t-SNE. Clusters were then identified using the density peak clustering option in Monocle. As previously, we removed clusters with less than 40 nuclei and then merged pairs of clusters with less than 10 differentially expressed genes (>1 natural log difference between clusters and expressed in >20% of cells in one of the two clusters). We also removed one putative doublet

cluster based on co-expression of VLMC markers (*Col1a1*) (*108*) and neuronal markers (*Meg3*) (*104*). This led to the identification of 44 different of cell types in the spinal cord.

Inferring spatial origin of spinal cord nuclei

To infer the spatial origin of each spinal cord cluster, we used annotated P4 ISH maps from the Allen Spinal Cord Atlas (*47*). Each gene has been manually annotated with 11 binary values to describe different expression patterns (Laminae 1-3, Laminae 4-6, Laminae 7-8, Laminae 9, Intermediolateral Column, Gray Matter, White Matter, Central Canal, Ventral-dorsal Midline in Gray Matter, Radially Arrayed in White Matter, and Vascular-like in Gray and White Matter).

We then used these data to create a composite map of differentially expressed genes for each cluster. For each cluster, the top 10 enriched genes (with spinal cord ISH data) were determined using differential gene expression described above. We then plotted the fraction of these 10 genes with expression in each laminae/region in the spinal cord (see Fig. S20 for all neuronal clusters).

Comparing the sensitivity of SPLiT-seq to droplet-based methods

To compare the sensitivity of SPLiT-seq to droplet-based approaches, we measured the number of UMIs and genes detected in mouse NIH/3T3 cells for SPLiT-seq, Drop-seq, and 10x Genomics (Chromium v2 chemistry) as a function of raw sequencing reads per cell. For Drop-seq, we analyzed the 100 STAMP dataset collected in Macosko et. al. (*6*) (SRA: SRR1748412). For 10x Genomics, we analyzed the 100 cell dataset available on their website (https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/hgmm_100).

The same pipeline was used to process each sample, with the only modifications made to account for the changes in cell barcode and UMI lengths. In the first step, reads with >1 base with

quality score less than phred 10 in the UMI were discarded. Reads were then aligned with STAR defaults to a combined mouse-human genome. We fixed cell barcodes with an edit distance of <=1 for all three methods. UMIs that were <=1 edit distance and corresponded to both the same cell barcode and gene were then collapsed. When we generated digital count matrices, we included intronic reads for 10x Genomics and SPLiT-seq, but excluded them from DropSeq because they led to a substantial increase in species impurity. We then subsampled each dataset between 5,000-50,000 raw reads per NIH/3T3 cell and recorded the number of UMIs as well as genes detected per cell (Fig. S3).

Detecting CDR3 regions in T-cells

To detect CDR3 regions present in SPLiT-seq and CleavR-SPLiT-seq libraries, reads from fastq files deriving from read 1 were first aligned using MiXCR (*109*) using default settings for RNA-seq data. Briefly, these steps included aligning reads, assembling partial reads, extending TCR alignments that lead to unique V and J genes, assembling clones, and finally exporting clones. Once clones are extracted, these clonotypes must be linked to the corresponding SPLiT-seq cell barcode present on the same read.

For fastq files where read 1 contributed to an exported clone, the read 2 counterpart was extracted (read with the cell barcode and UMIs). Next, these cell barcodes were cross-referenced with barcodes determined to be considered a cell based on cDNA data. This generates a table matching each aligned read containing a CDR3 region to a cell barcode. Alignments were then grouped by cell barcode. For each aligned read, MiXCR also outputs an associated alignment confidence score. In some cases, cell barcodes had more than one TRA or TRB CDR3 alignment. In these cases, a filter was applied where the highest scoring alignment had encompass a score twice that of the sum of any other alignments. When false, all alignments for that cell were

removed and not considered in future analysis. After this step, percentages of cells containing TRA CDR3 and TRB CDR3 alignments were determined by dividing the total number of cells containing that CDR3 alignment over the total number of cells present in the library, as determined by cDNA reads.

Generating SSO Target Finder Scores

Scores for exon skipping and inclusion for every possible splice switching oligonucleotide within a given exon are generated by first calculating significance scores for every individual base in the exon. Significance scores are calculated by using HAL's model to predict exon skipping (*74*). The input for HAL is the wildtype exon sequence, the same exon sequence with the variant of interest introduced, and the percent the wild type sequence is usually spliced in (PSI). HAL outputs a predicted change in spliced in ratio (DPSI) based on the variant introduced. To produce significance scores, this model is run four times for each base, representing a variants where the wild type base is either substituted into one of the three other bases or deleted. These resulting four DPSI scores are averaged to acquire the significance score. The goal here is to determine the overall impact of the wild type base present here rather than the impact of a specific change (i.e. specific base substitution).

Once significance scores for each base are calculated, the SSO length can be initialized. If not initialized, the default will be to scan for all possible SSOs within range of 14-20 bases in length. For the example comparing SSOs tested from Hua *et al* with the SMN2 exon 7 region (*87*), the SSO target finder was set to only screen SSOs with the same length tested in the publication to make a proper comparison (15 bases). Once the SSO length is initialized, an oligowalk is performed where a score is calculated for each possible SSO in the exon with those lengths. In practice, the significance scores of all bases being covered by a given SSO are added together to

form a score for each SSO. All SSOs are then ranked based on the SSO score. Lowest scores indicate SSOs predicated to promote highest exon exclusion, while highest scores indicate SSOs predicated to promote highest exon inclusion. The theory behind this is that bases that have very negative DPSI values indicate that these bases are critical to the inclusion of the exon. Therefore, applying an SSO to essentially block functionality of that same sequence should lead to increased probability of exon exclusion. This same principle can be applied to bases that contain high positive DPSI values for exon inclusion.

Supplementary Figures

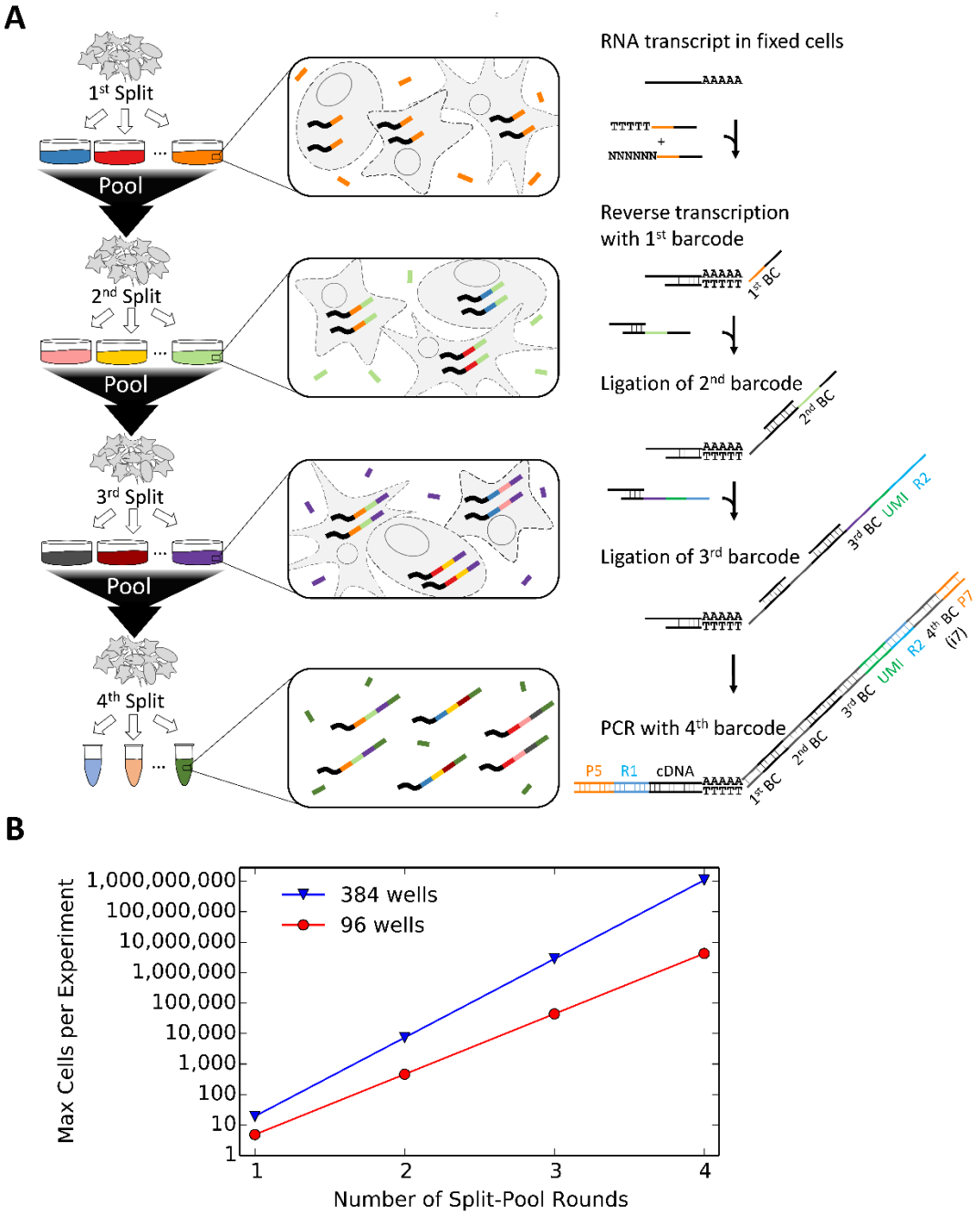


Fig. S1. Molecular diagram and exponential scalability of SPLiT-seq. (A) Fixed and permeabilized cells are randomly split into wells that each contain reverse transcription primers with a well-specific barcode. *In situ* reverse transcription converts RNA to cDNA while appending the well-specific barcode. Cells are then pooled and again randomly split into a second set of wells, each containing a unique well-specific barcode. These barcodes are hybridized and ligated to the 5'-end of the barcoded reverse transcription primer to add a second round of barcoding. The cells

are pooled back together and a subsequent split-ligate-pool round can be performed. After the last round of ligation, cDNA molecules contain a cell-specific combination of barcodes, a unique molecular identifier, and a universal PCR handle on the 5'-end. A fourth barcoding round is performed during the PCR step of library preparation. **(B)** Exponential scalability of SPLiT-seq with number of split-pool rounds. The maximum number of cells was calculated with the assumption that the number of barcode combinations must be twenty times greater than the number of cells.

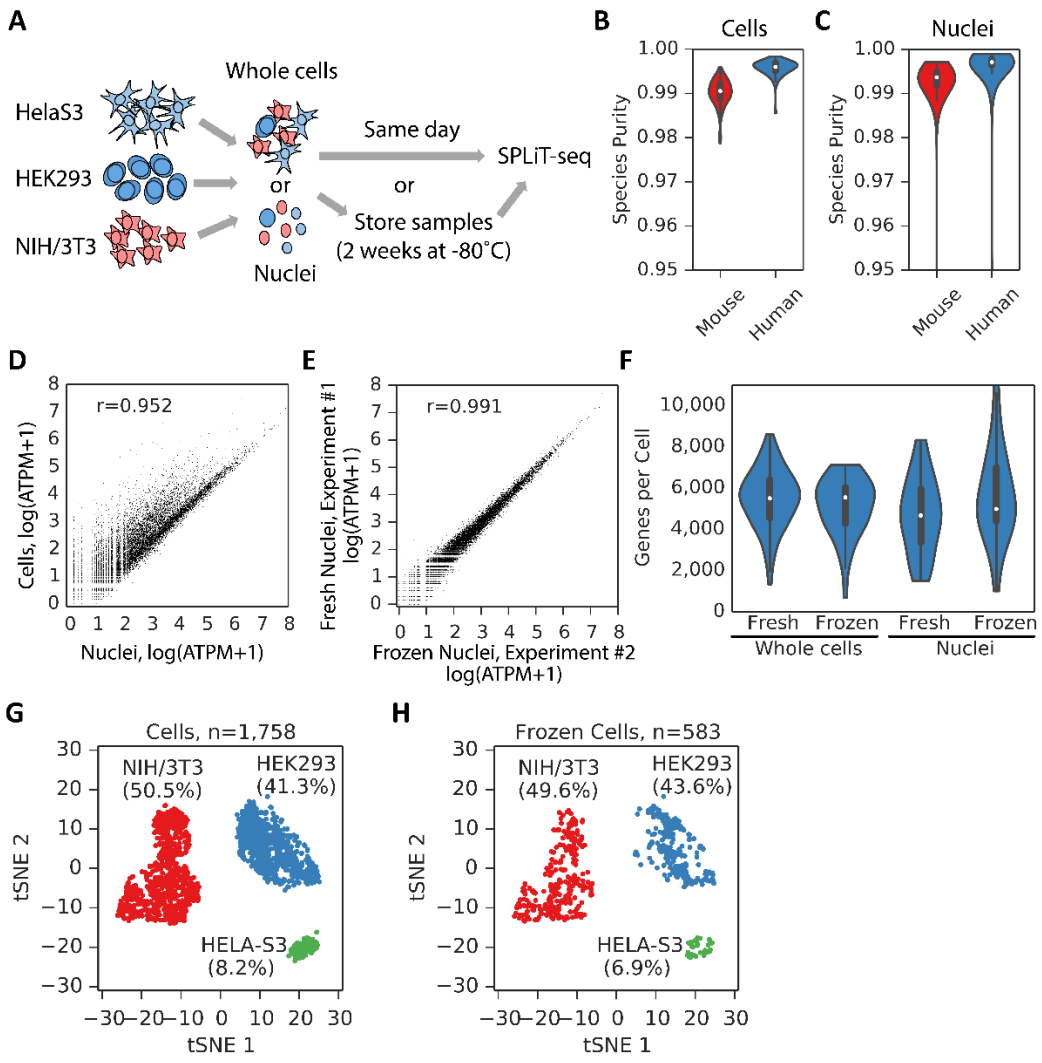


Fig S2. SPLiT-seq performance evaluation: species purity, gene expression correlation, gene detection, and cell preservation. (A) Cells or nuclei of different origin (e.g. mouse and human cell lines) were mixed and profiled with SPLiT-seq. Cells were either processed immediately or after two weeks at -80˚C. **(B)** Fraction of reads mapping to the correct species for mouse and human cells. **(C)** Fraction of reads mapping to the correct species for mouse and human nuclei. **(D)** Gene expression in nuclei and whole cells is highly correlated (Pearson-r: 0.952). Average gene expression across all cells (log average transcripts per million) is plotted for each experiment. **(E)** Gene expression from frozen and stored nuclei is highly correlated to nuclei processed immediately (Pearson-r: 0.991). **(F)** Gene counts from mixing experiments performed with fresh and frozen whole cells and nuclei. Median gene counts for fresh cells: 5,498; frozen cells: 5,540; nuclei: 4,663; frozen nuclei: 4,982**. (G)** Storing cells at -80C for two weeks does not affect cell type identification. Immediately processed cells as well as frozen and stored cells were clustered together using t-SNE. Immediately processed cells and frozen cells cluster according to cell type rather than batch/processing method. Similar proportions of cells in each cluster are maintained for frozen cells.

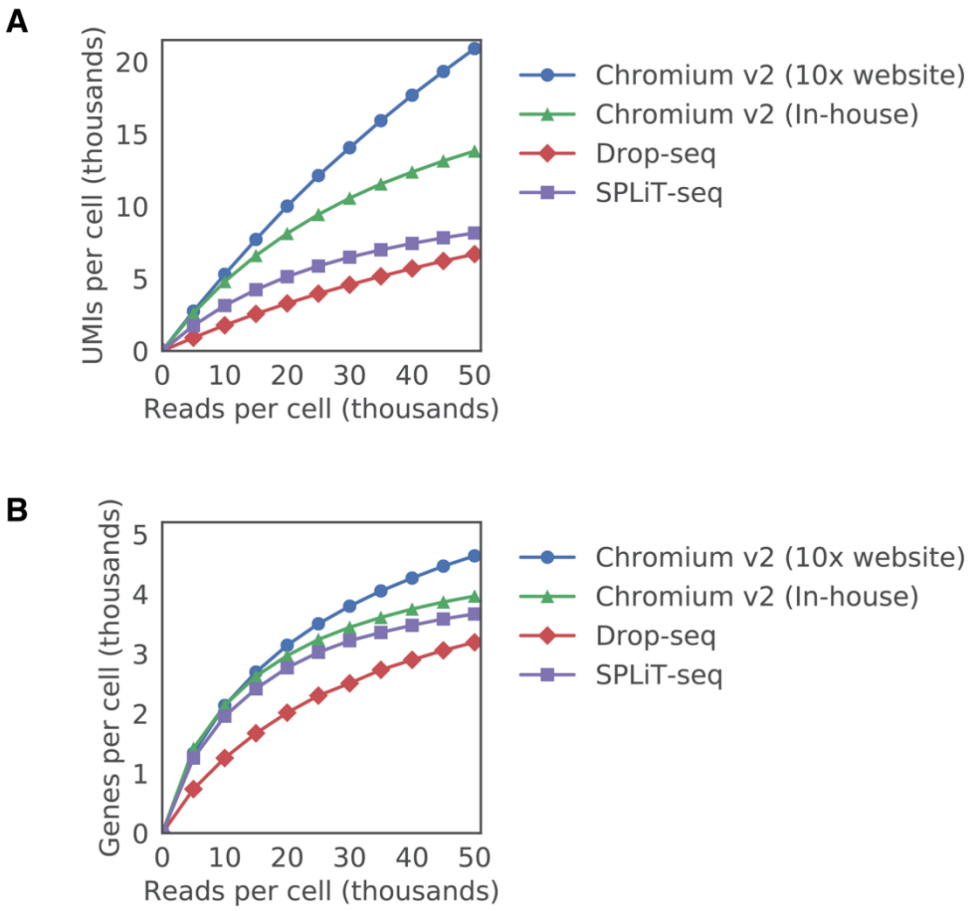


Fig. S3. Downsampling comparison of SPLiT-seq to other scRNA-seq methods. Median UMIs **(A)** or genes **(B)** detected per mouse cell (NIH/3T3) are shown as a function of raw sequencing reads. The reads for all cells were down-sampled from 50,000 to 5,000 in increments of 5,000. We compared 10x Genomics data collected in-house (at the Allen Institute for Brain Science) as well as the best available dataset on the 10x Genomics website with respect to detected UMIs per cell. Drop-seq data was taken from Macosko *et al*, while data used for SPLiT-seq was taken from the species-mixing experiment presented in Fig. 1.

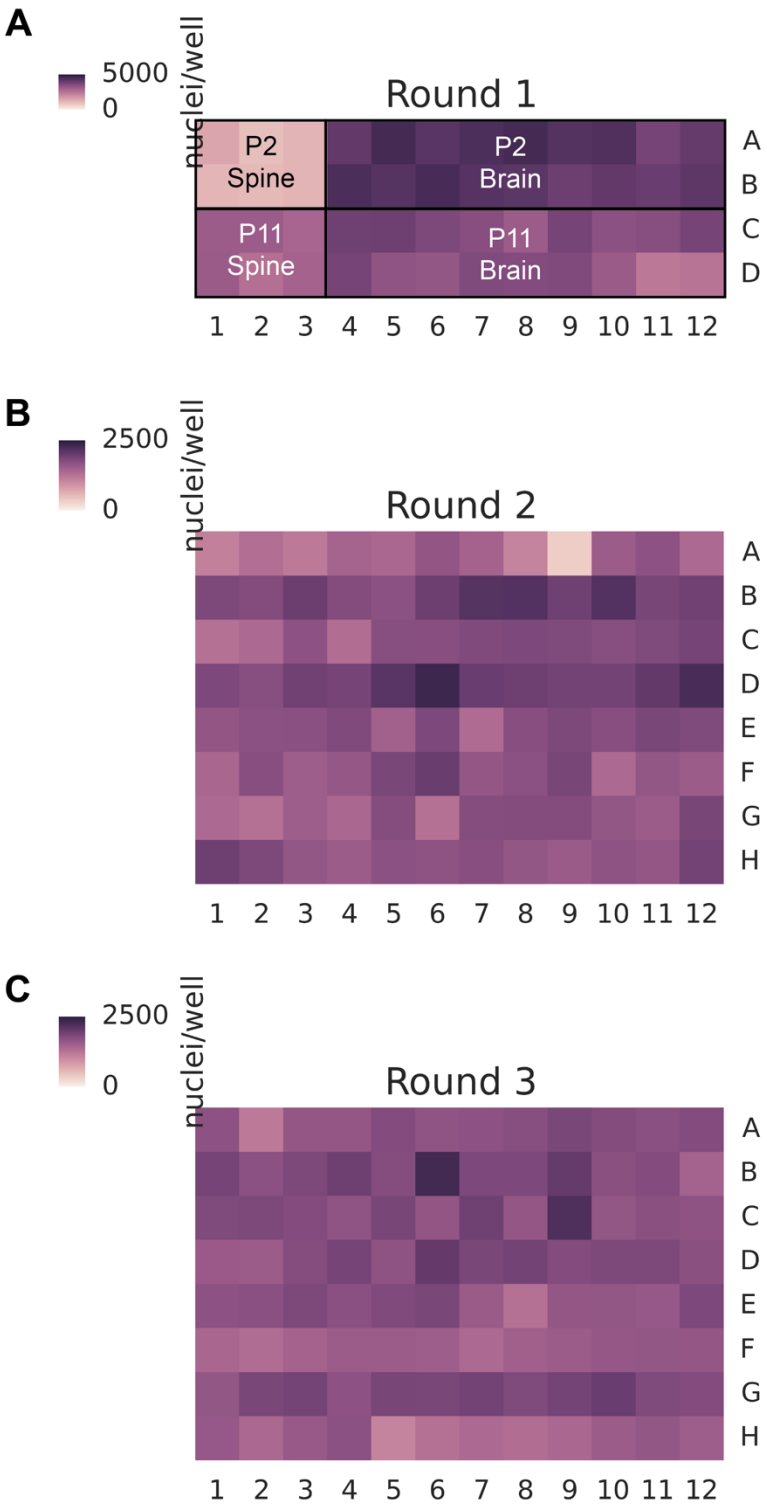


Fig. S4. Number of nuclei in each well during three rounds of barcoding. Despite pipetting cells by hand, most wells contain approximately equal numbers of nuclei. Dissociation of the P2 spinal cord resulted in fewer cells than the other samples, explaining the lower number of nuclei in the corresponding first round wells.

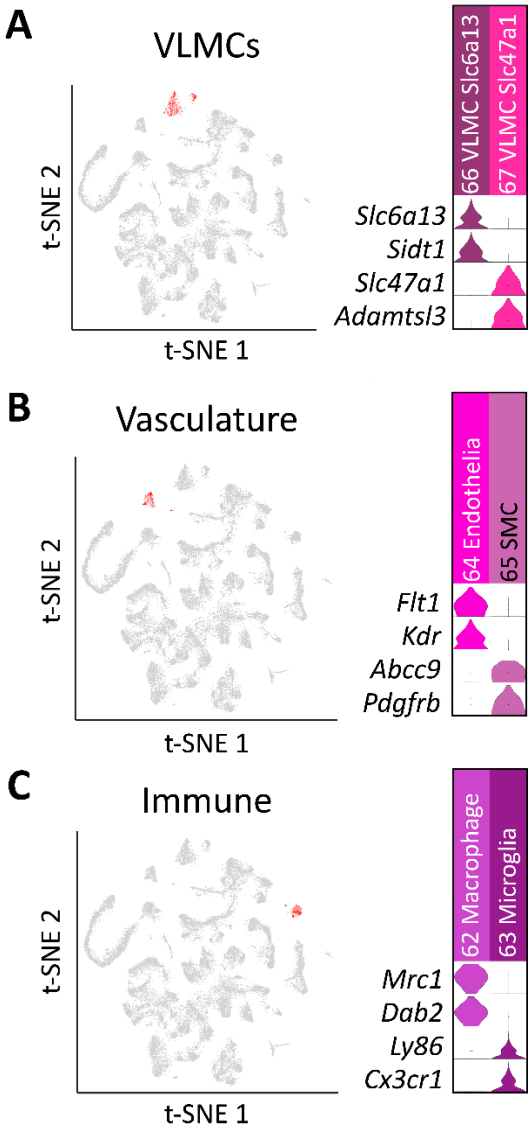


Fig. S5. Differences in VLMC, vasculature, and immune cell types. (A) During postnatal development, one VLMC subtype was found to differentially express *Slc6a13* and *Sidt1* whereas another subtype was found to differentially express *Slc47a1* and *Adamtsl3*. **(B)** Endothelia were found to differentially express *Flt1* and *Kdr* whereas smooth muscle cells were found to differentially express *Abcc9* and *Pdgfrb*. **(C)** Macrophages differentially express *Mrc1* and *Dab2* whereas microglia differentially express *Ly86* and *Cx3cr1*.

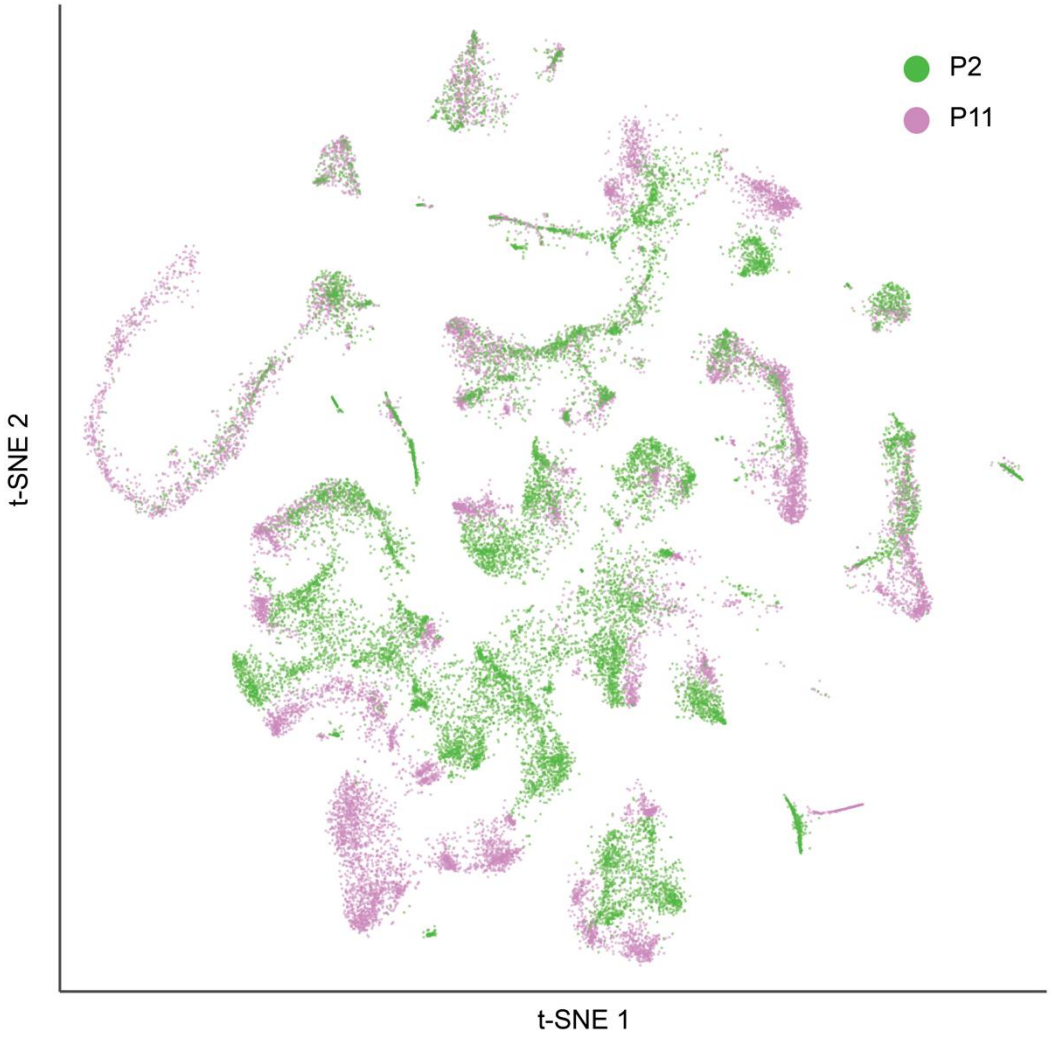


Fig. S6. Distribution of P2 and P11 transcriptomes projected with t-SNE.

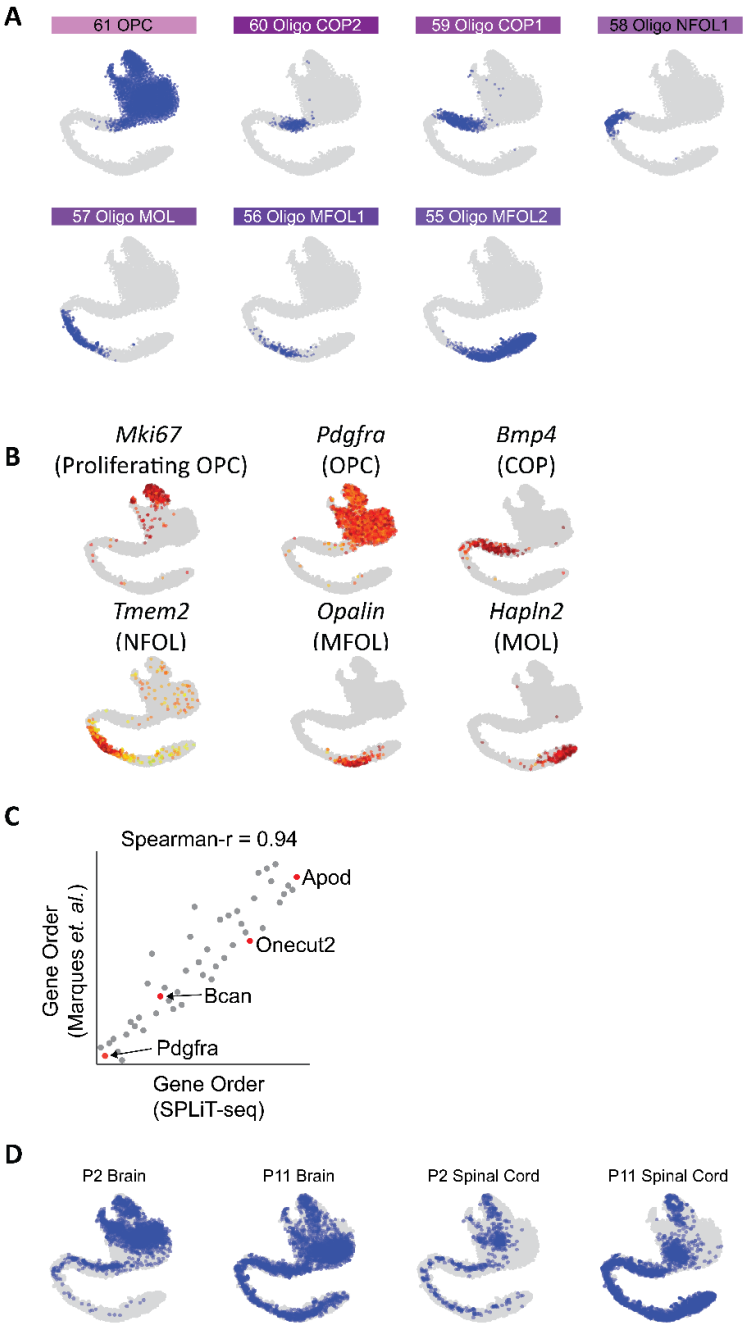


Fig. S7. Oligodendrocyte lineage. (A) Single-nucleus transcriptomes from seven clusters within the oligodendrocyte lineage were re-embedded with t-SNE. **(B)** Gene markers overlaid on the re-embedded t-SNE show proliferative markers like *Mki67* on one end with mature markers like *Hapln2* on the other end **(C)** Comparison of gene ordering between our oligodendrocyte lineage and that in Marques *et. al*. **(D)** Distribution of P2/P11 brain and spinal cord single-nucleus transcriptomes within the oligodendrocyte lineage.

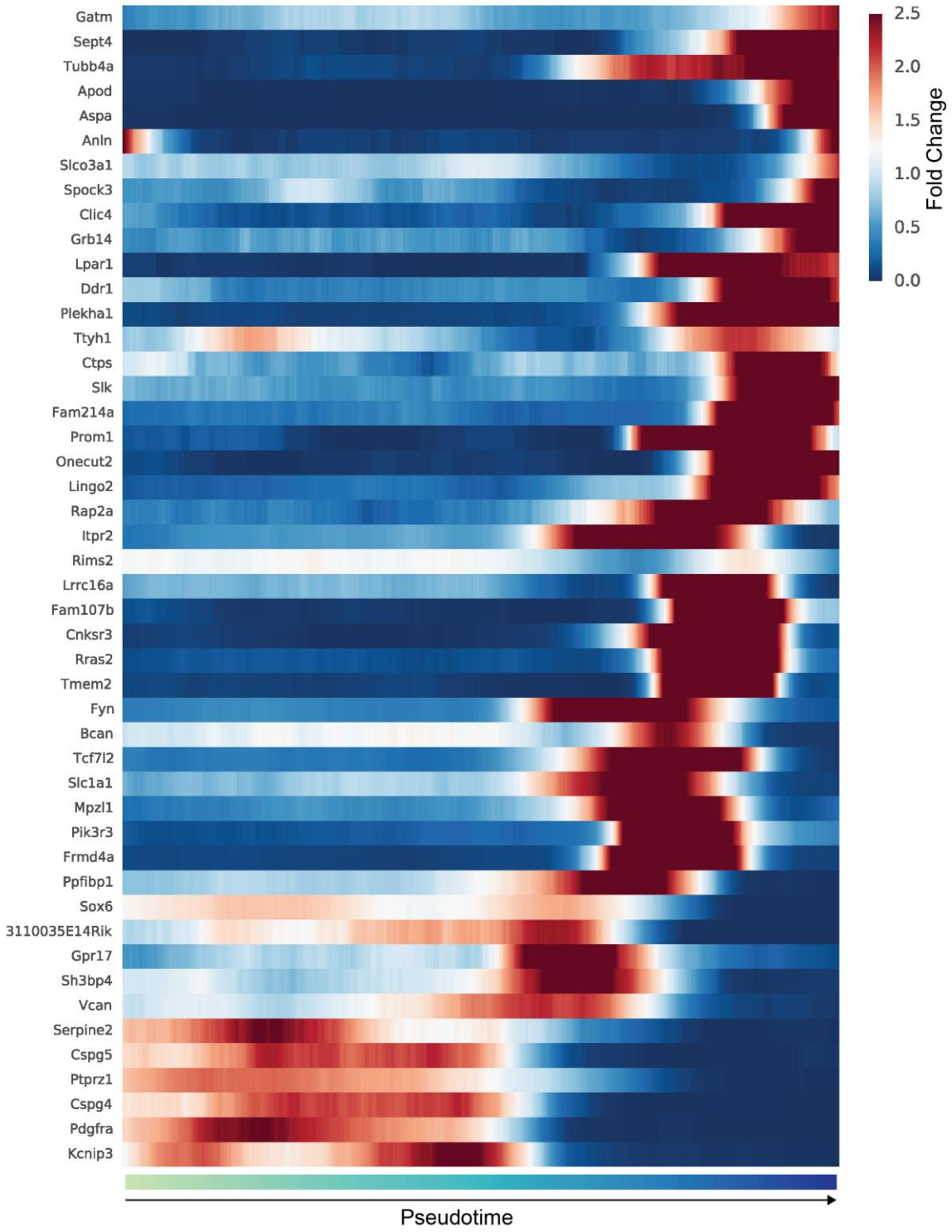


Fig. S8. Gene expression in oligodendrocyte lineage. Genes are chosen from Marques *et. al.*(*9*). Fold change is calculated relative to mean gene expression in the entire lineage.

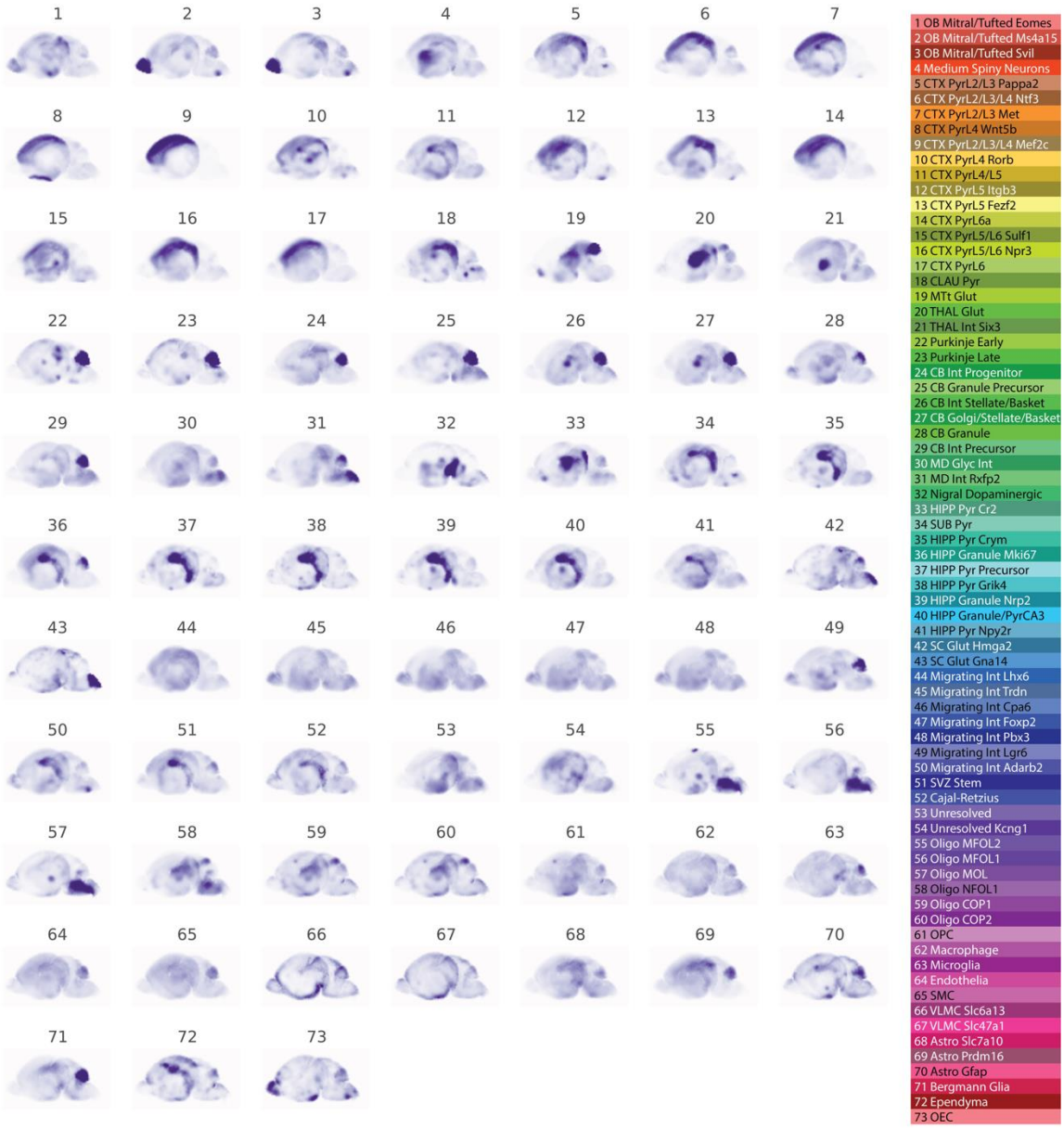


Fig. S9. Composite P4 ISH maps generated from the Allen Developing Brain Atlas. For each cluster, we selected the 5 genes most enriched in that cluster. We then averaged P4 ISH data for these genes and plotted the cumulative ISH signal across sagittal slices.

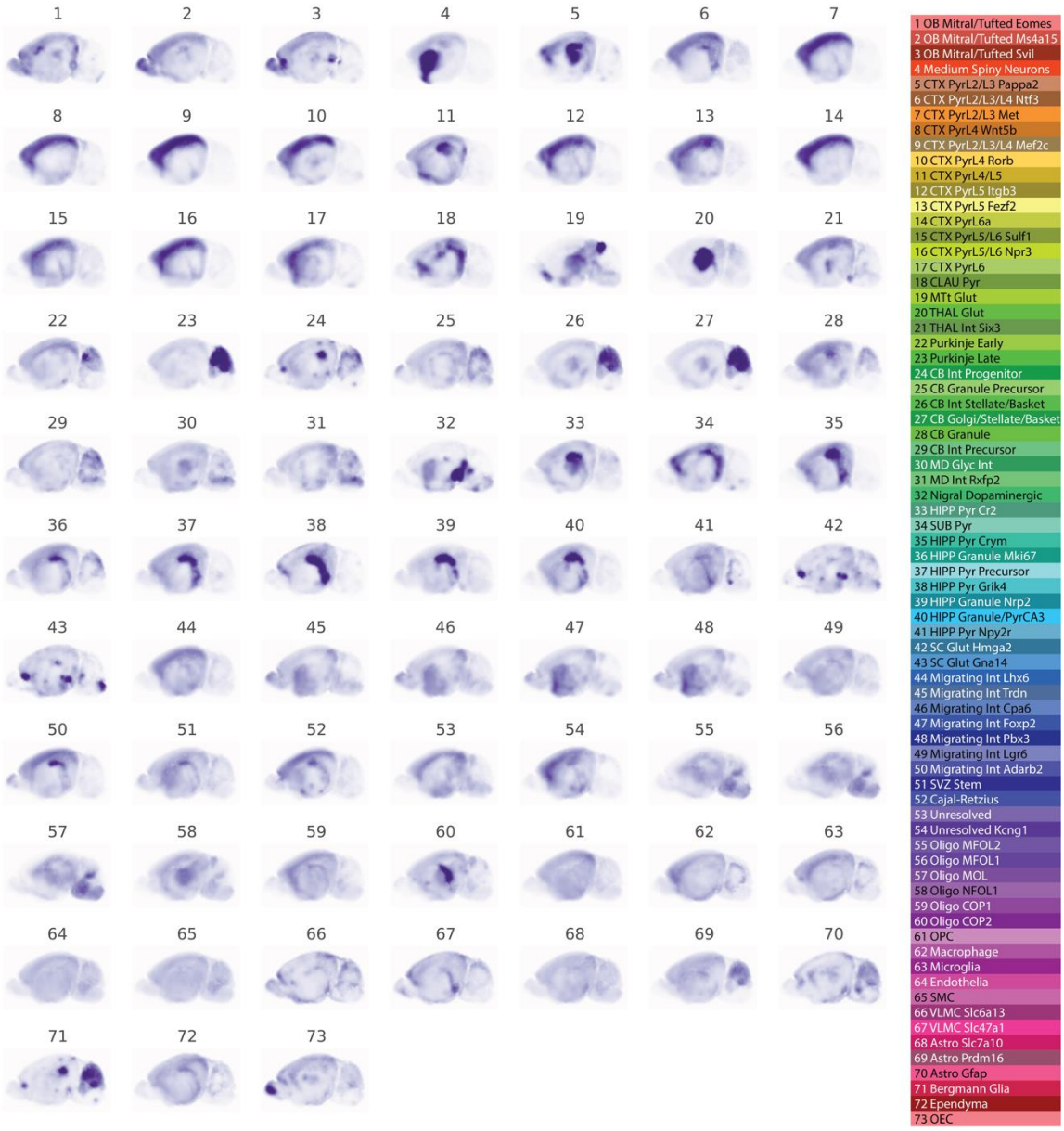


Fig. S10. Composite P14 ISH maps generated from the Allen Developing Brain Atlas. For each cluster, we selected the 5 genes most enriched in that cluster. We then averaged P14 ISH data for these genes and plotted the cumulative ISH signal across sagittal slices.

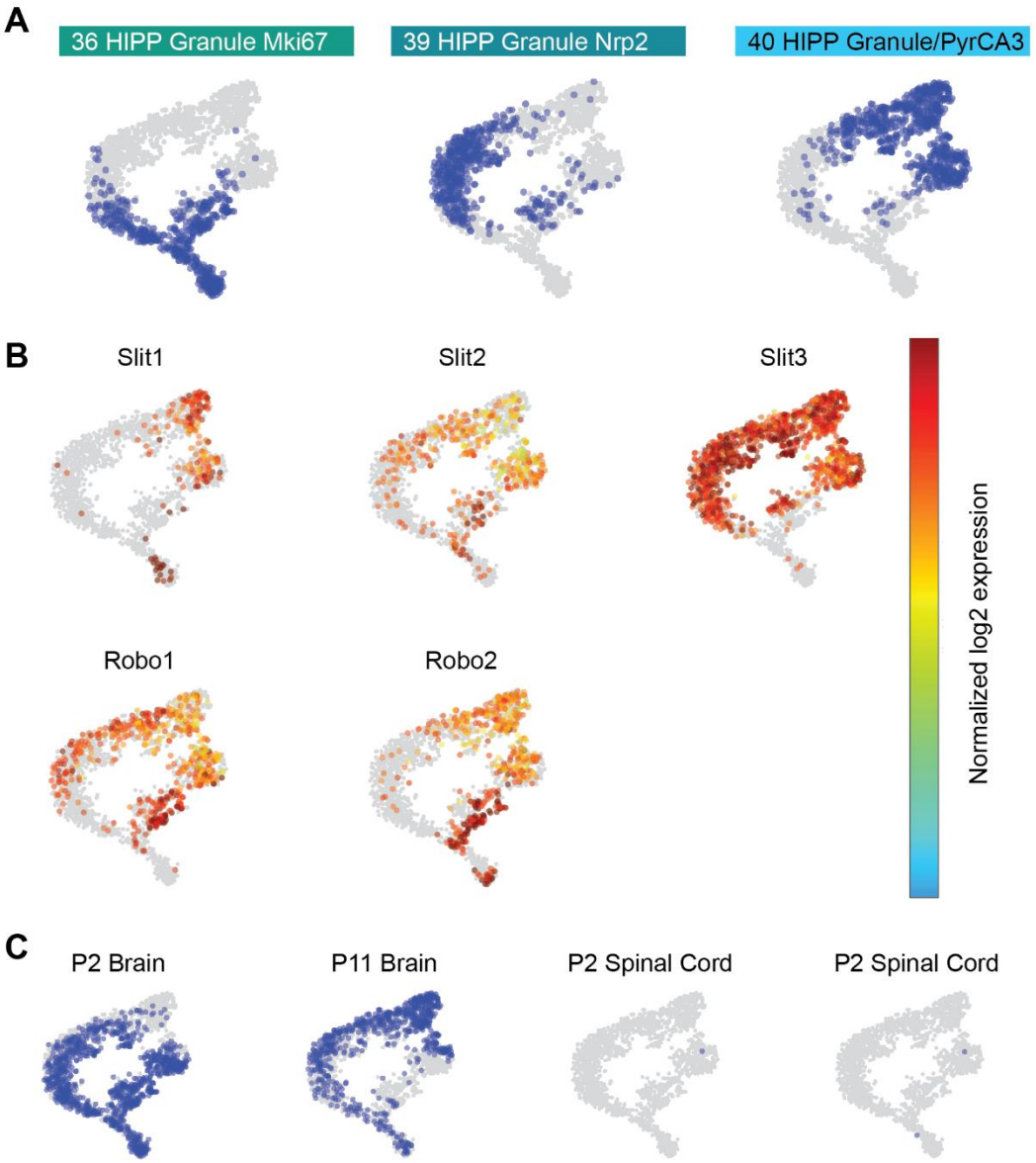


Fig. S11. Branching hippocampal neuronal lineage. (A) Three hippocampal clusters were re-embedded with t-SNE. The original clusters are overlaid over the resulting t-SNE. **(B)** Dynamics of *Slit1/2/3* and *Robo1/2* across pseudotime. **(C)** Distribution of P2/P11 brain and spinal cord single-nucleus transcriptomes within the lineage.

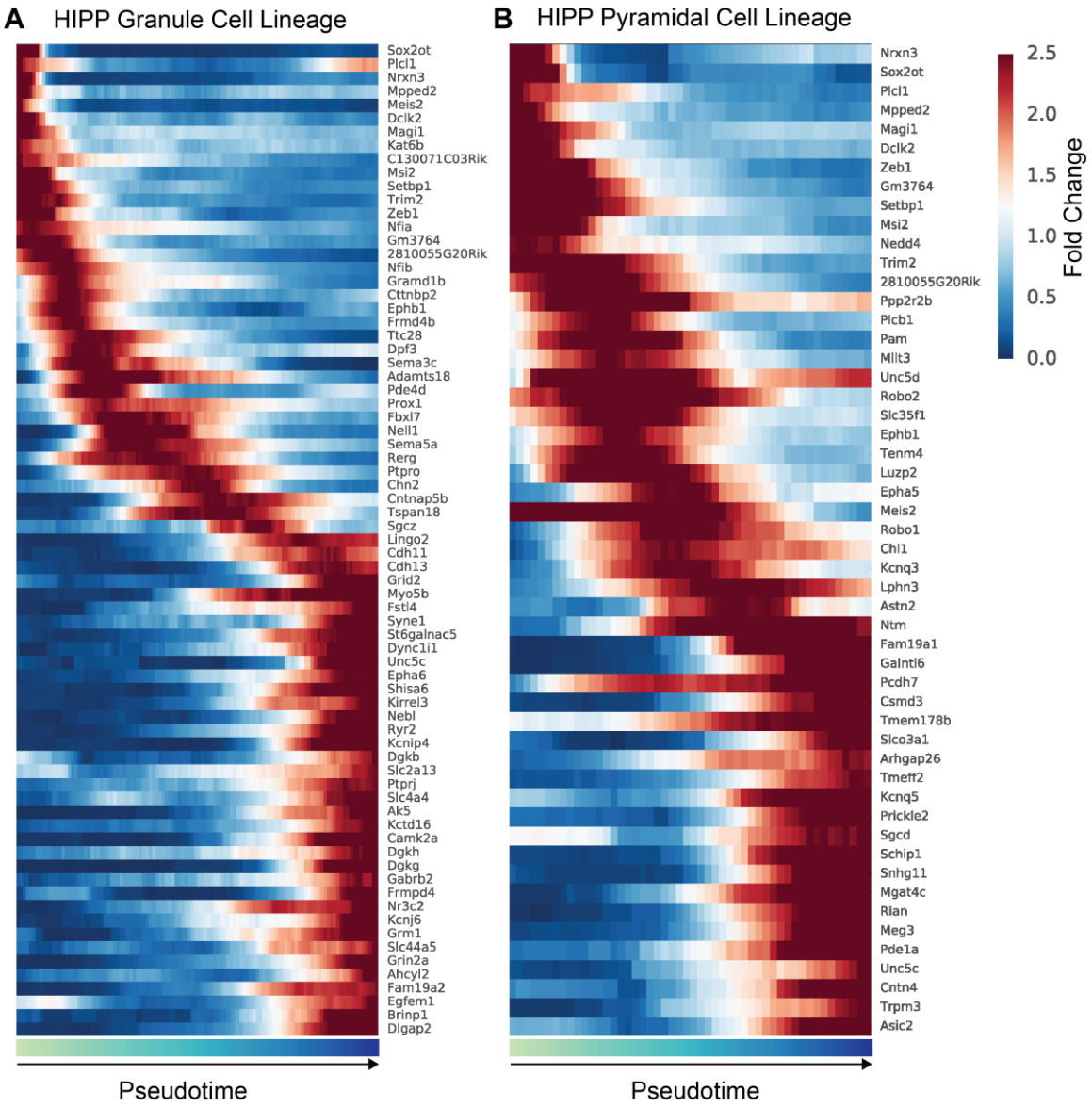


Fig. S12. Gene expression in branching hippocampal neuronal lineage. (A) Genes with differential expression across pseudotime in the granule cell lineage. **(B)** Genes with differential expression across pesudotime in the pyramidal cell lineage. Fold change is calculated relative to mean gene expression in the entire lineage.

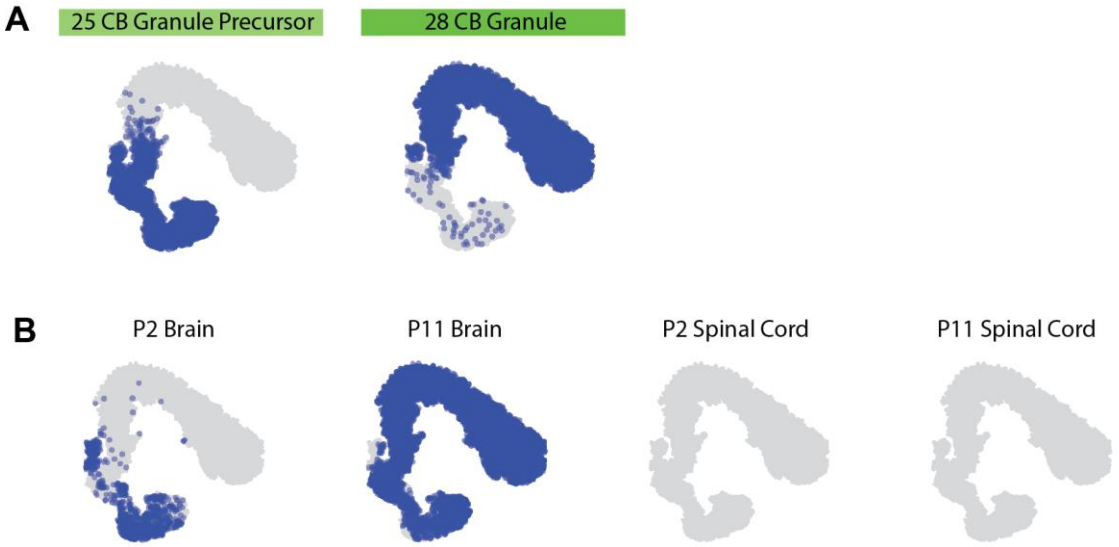


Fig. S13. (A) Cerebellar granule cell lineage. Transcriptomes from two cerebellar granule clusters were re-embedded with t-SNE. **(B)** Distribution of P2/P11 brain and spinal cord cells within the lineage.

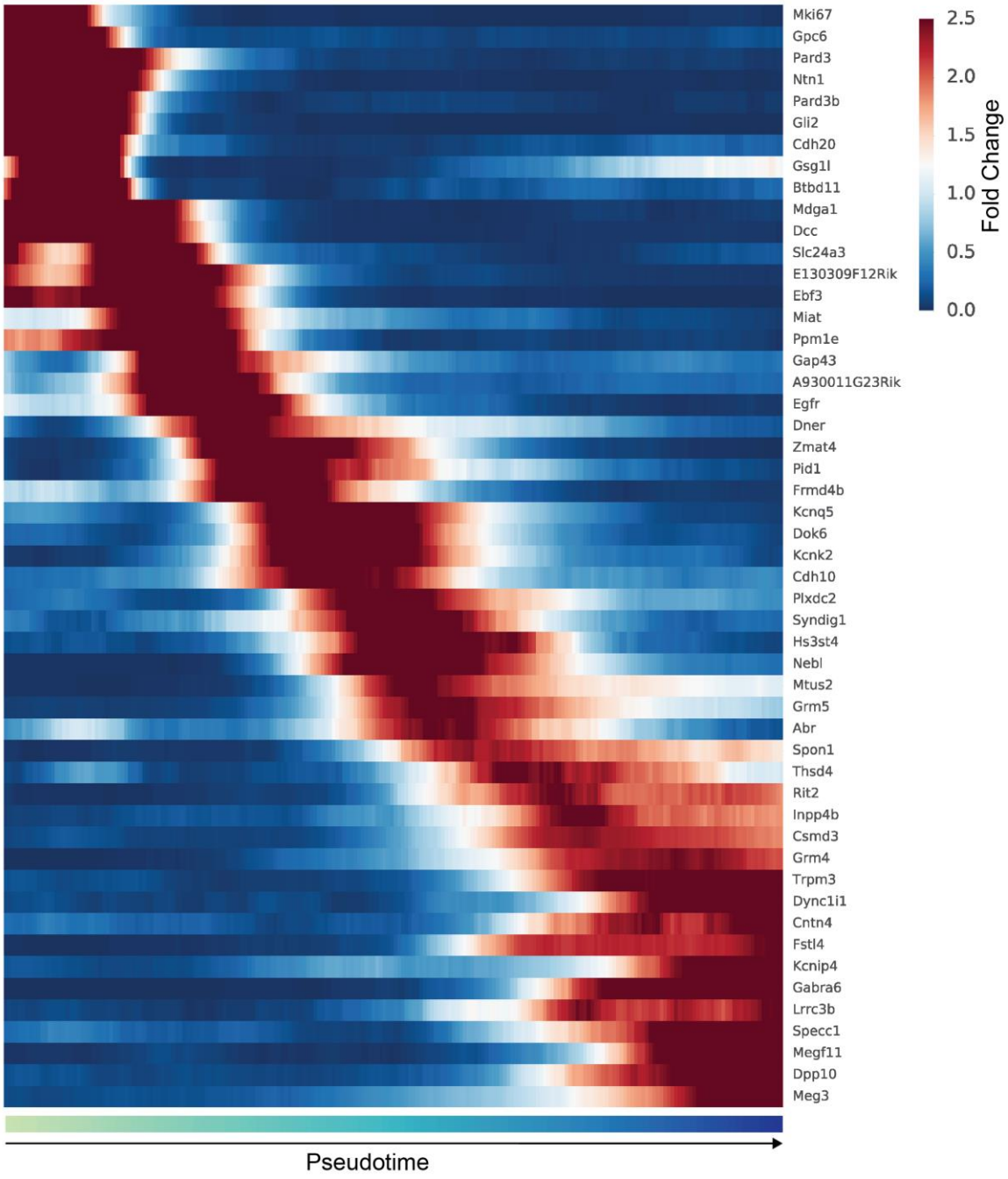


Fig. S14. Genes with differential expression across pseudotime in the cerebellar granule cell lineage.

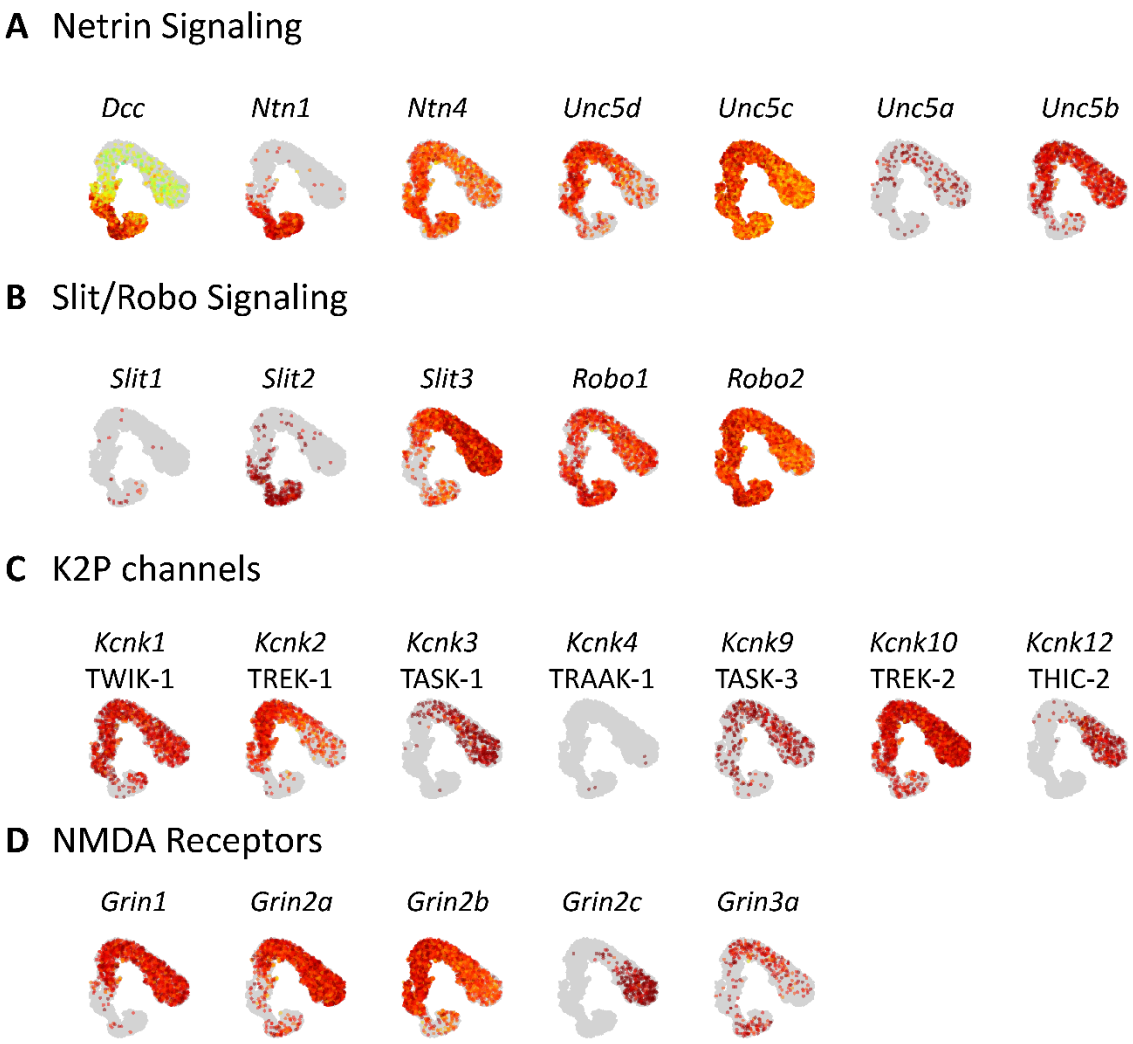


Fig. S15. Pathways relevant to cerebellar granule cell migration and development. Cerebellar granule cell lineage t-SNE overlaid with expression of genes contributing to **(A)** netrin signaling, **(B)** *Slit/Robo* signaling, **(C)** two-pore domain potassium (K2P) channels, and **(D)** N-methyl-D-aspartate (NMDA) receptors

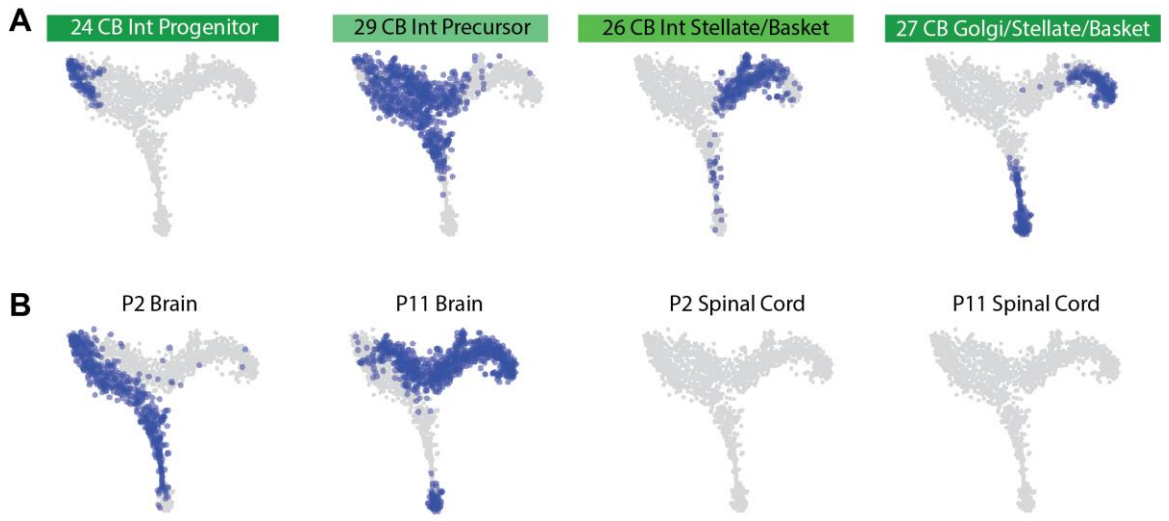


Fig. S16. (A) Cerebellar interneuron cell lineage. Four clusters were re-embedded with t-SNE. **(B)** Distribution of P2/P11 brain and spinal cord cells within the lineage.

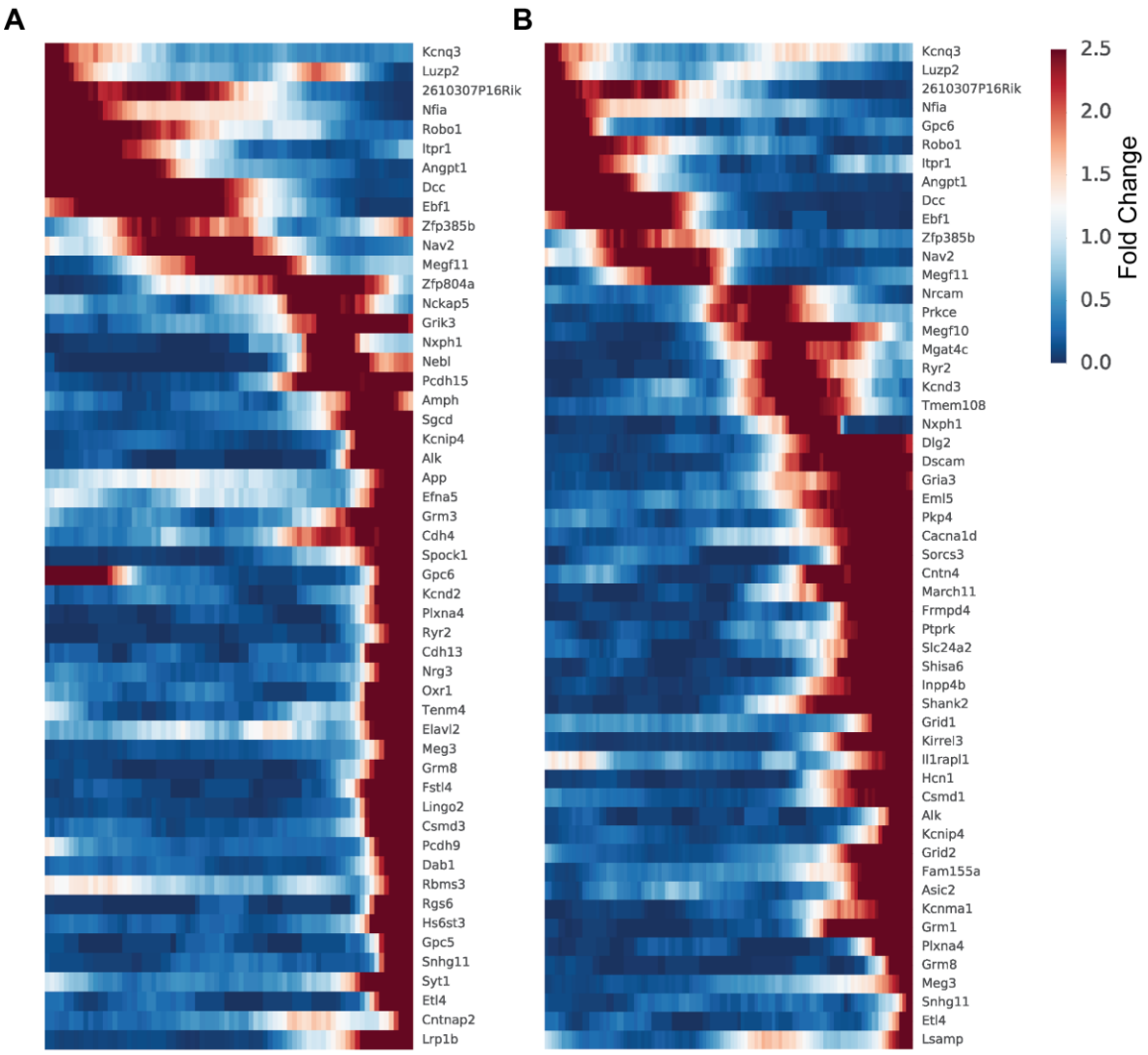


Fig. S17. Gene expression in branching cerebellar interneuron lineage. (A) Genes with differential expression across pseudotime in the stellate/basket cell lineage. **(B)** Genes with differential expression across pesudotime in the Golgi cell lineage.

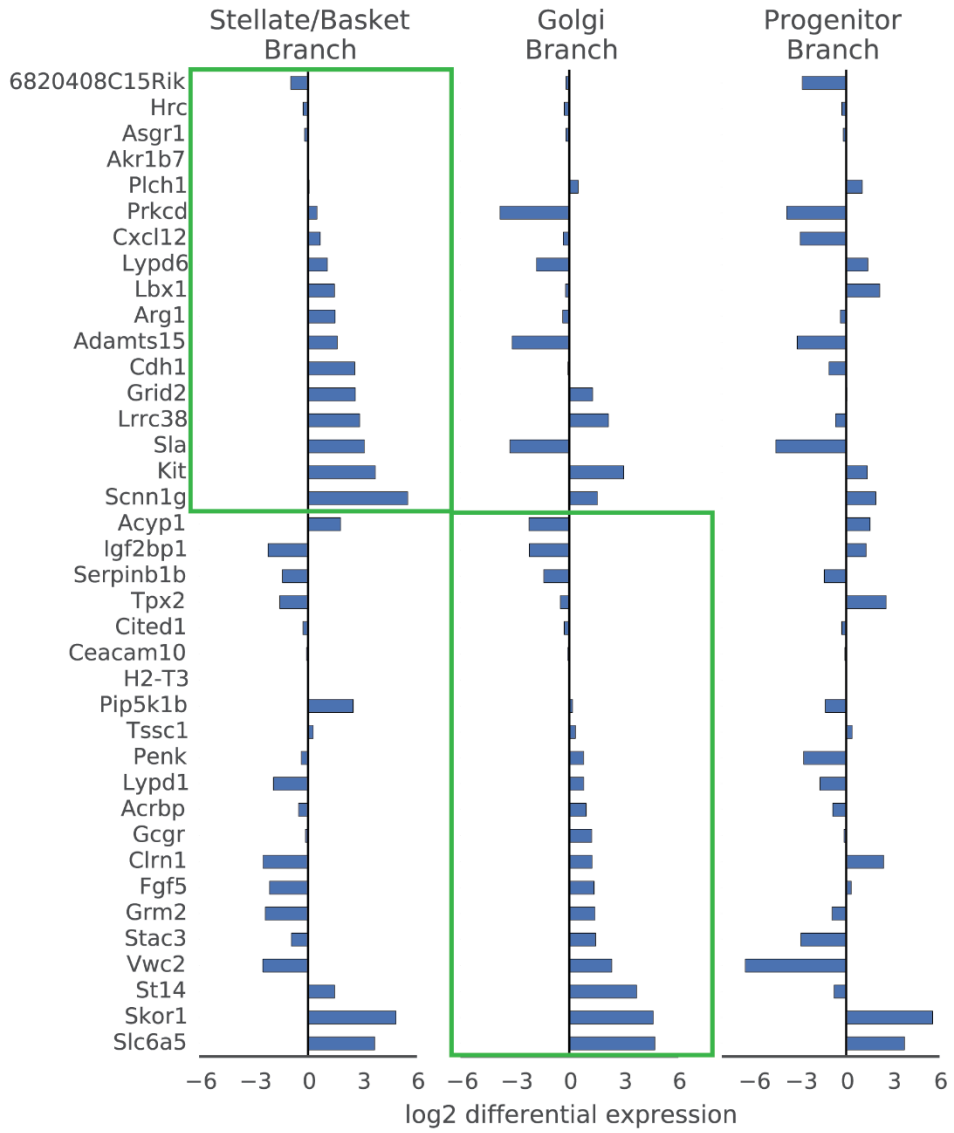


Fig. S18. Validating stellate/basket and Golgi cell identities of cerebellar interneuron lineage branches. Previously characterized marker genes of Golgi and stellate/basket cells (*55*) are plotted for each branch of the cerebellar interneuron lineage.

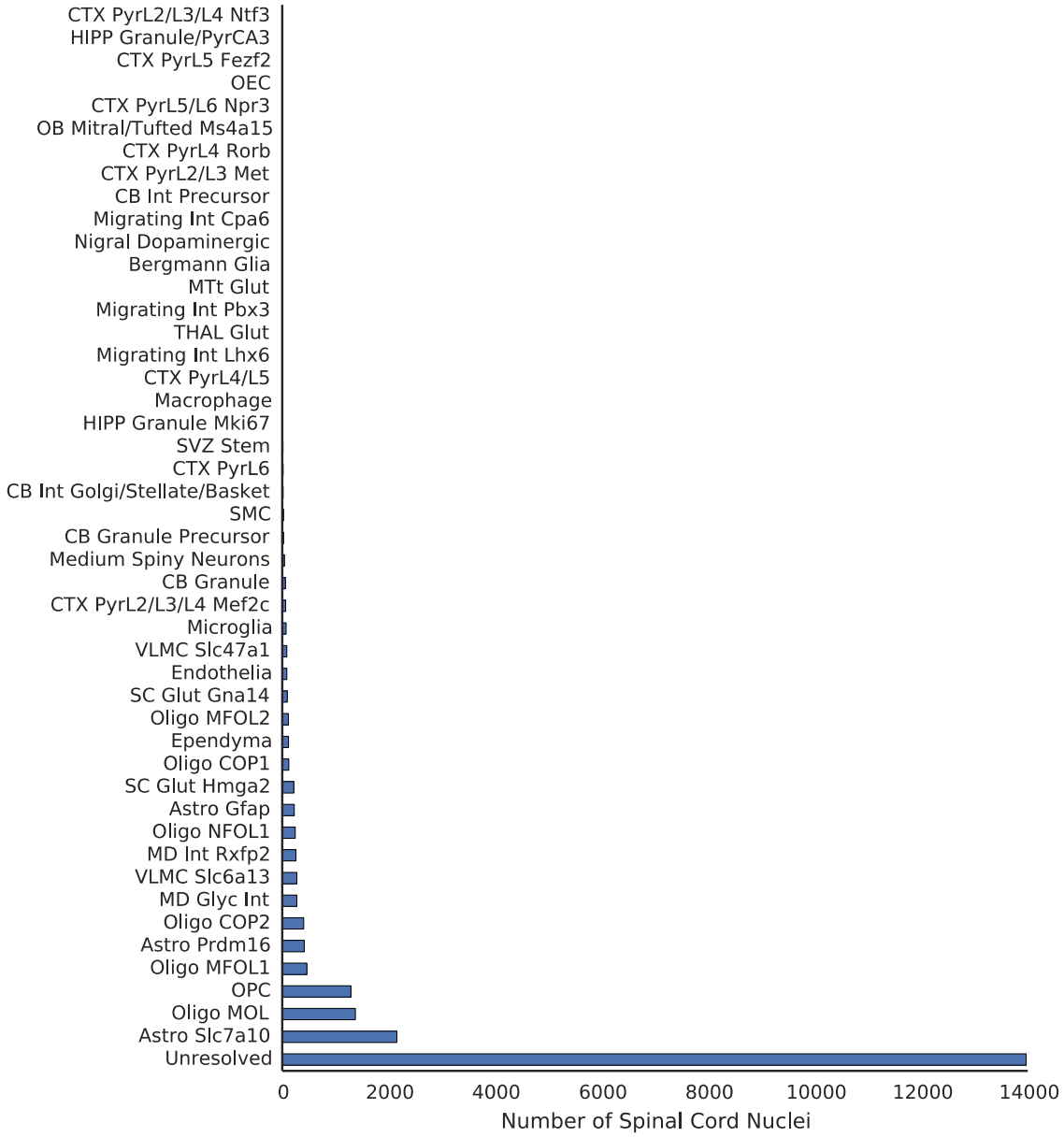


Fig. S19. Number of nuclei in each cluster from the spinal cord. Over 60% of spinal cord nuclei clustered into the unresolved cluster, leading us to re-cluster the spinal cord nuclei without brain nuclei.

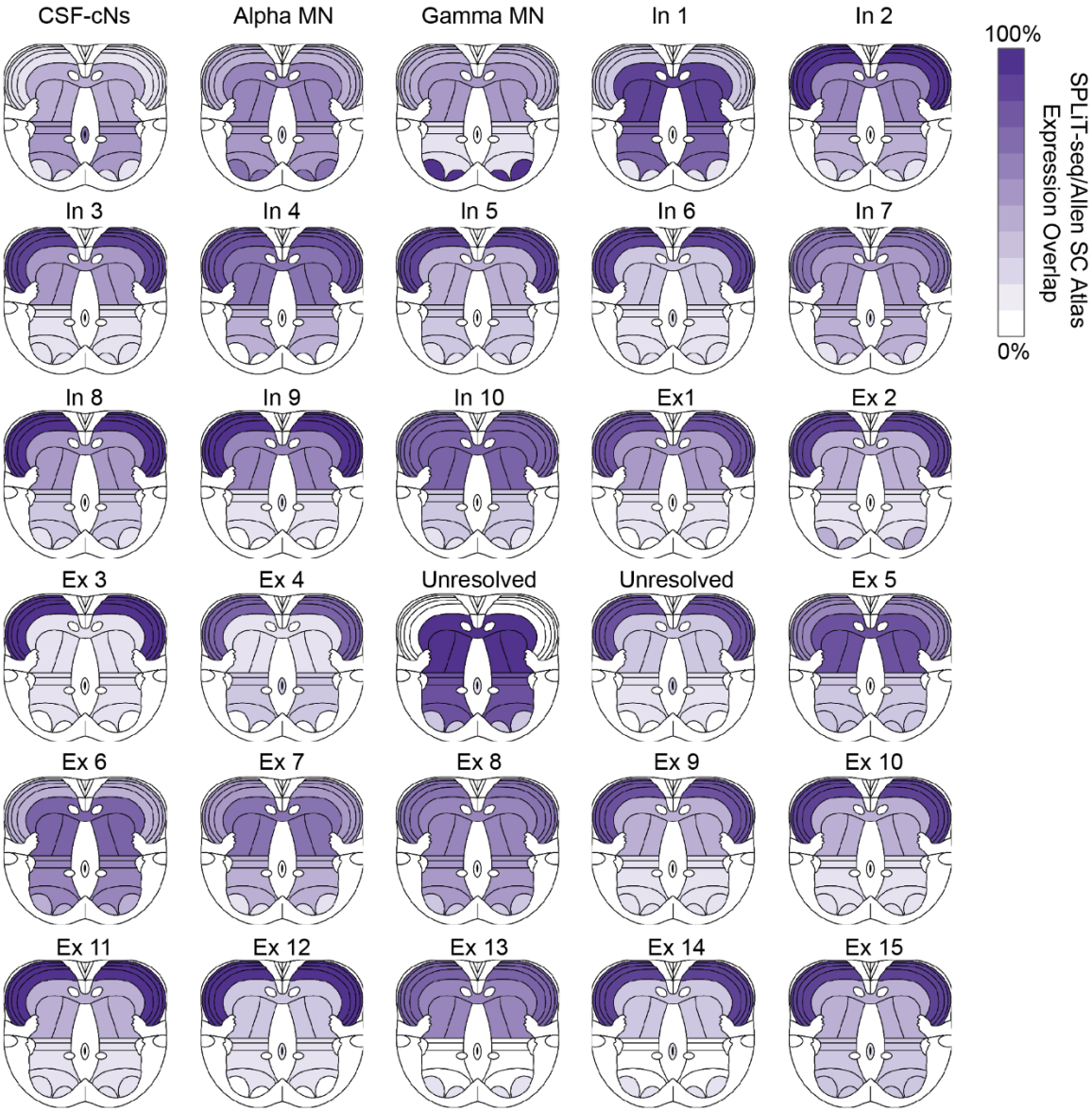


Fig. S20. Inferred spatial origin of all neuronal clusters within the spinal cord.

Inferred spatial origin of neuronal clusters within the spinal cord. We analyzed the Allen Spinal Cord Atlas expression patterns of the top ten enriched genes in each cluster. Dark purple indicates expression of all ten genes in the given region, while white indicates none of the ten genes were expressed in the given region.

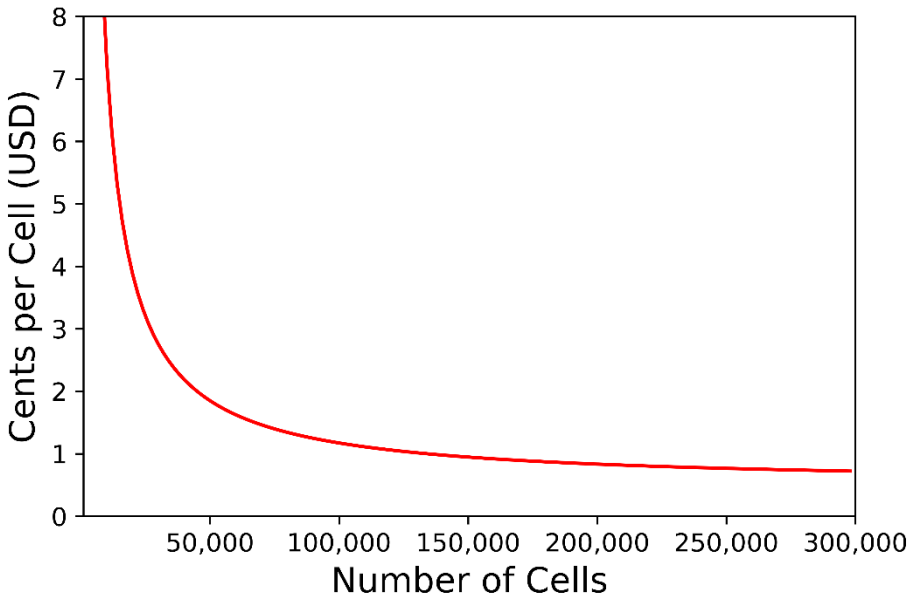


Fig. S21. Cost of library preparation per cell for SPLiT-seq. As more cells are processed, costs drop below 1 cent per cell, making SPLiT-seq a cost-effective platform to profile large numbers of cells. This analysis does not include Illumina sequencing cost.

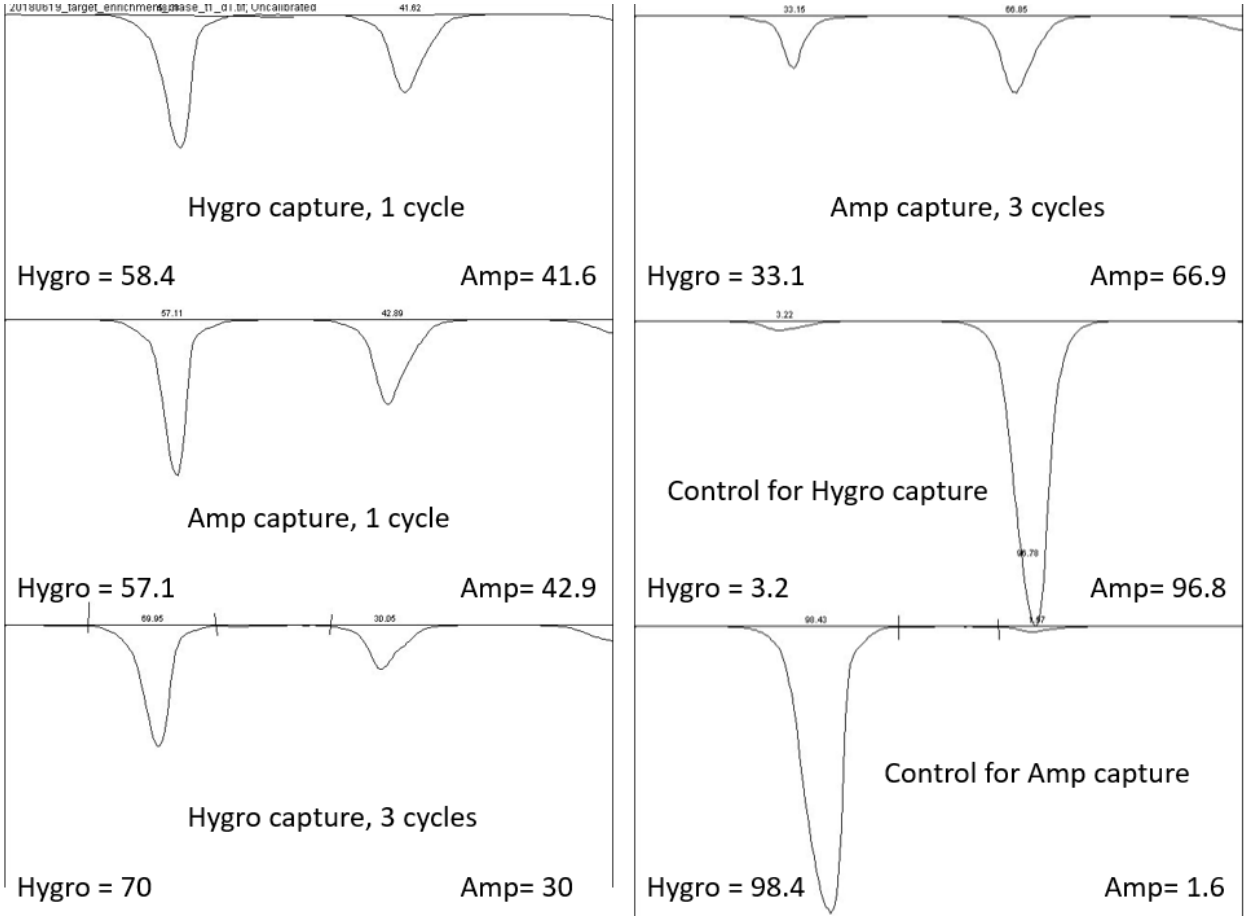
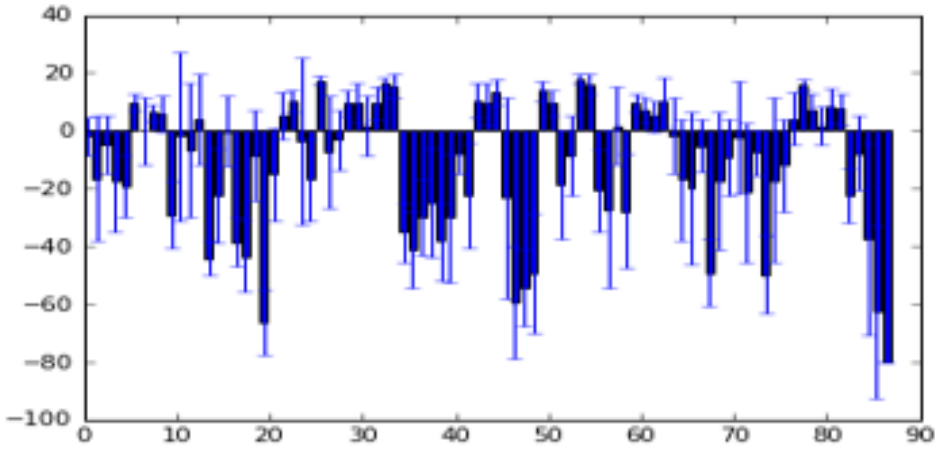


Fig. S22. CleavR fold enrichment calculation using ImageJ. A gel was run on the amplified products from the Amp/Hygro strand experiment to calculate fold enrichment, as shown in fig. 3.4. To calculate fold enrichment, specific weights of each of the two bands (Amp or Hygro strand) is needed. ImageJ was used to quantify the signal of each band for the six lanes corresponding to the experiment described in fig 3.4. Quantification for each lane is displayed here, with the percent of each band contributing to the total signal in the lane. In all six lanes, the additional of signal from both lanes equaled 100%, indicating little to no background signal.

A



B

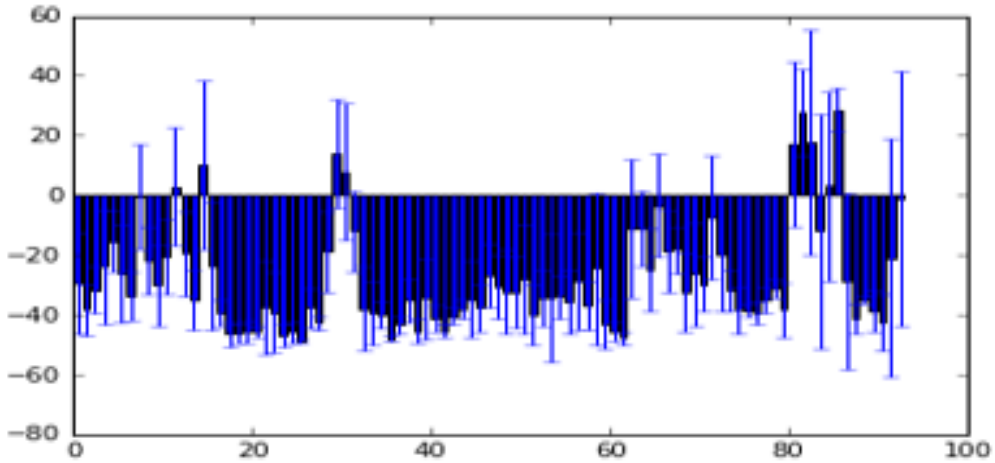


Fig. S23. SSO target finder head example heat maps. Example of other exon heat maps generated by SSO target finder looking at CFTR exon 12 (A) and (B) MAPT exon 10.

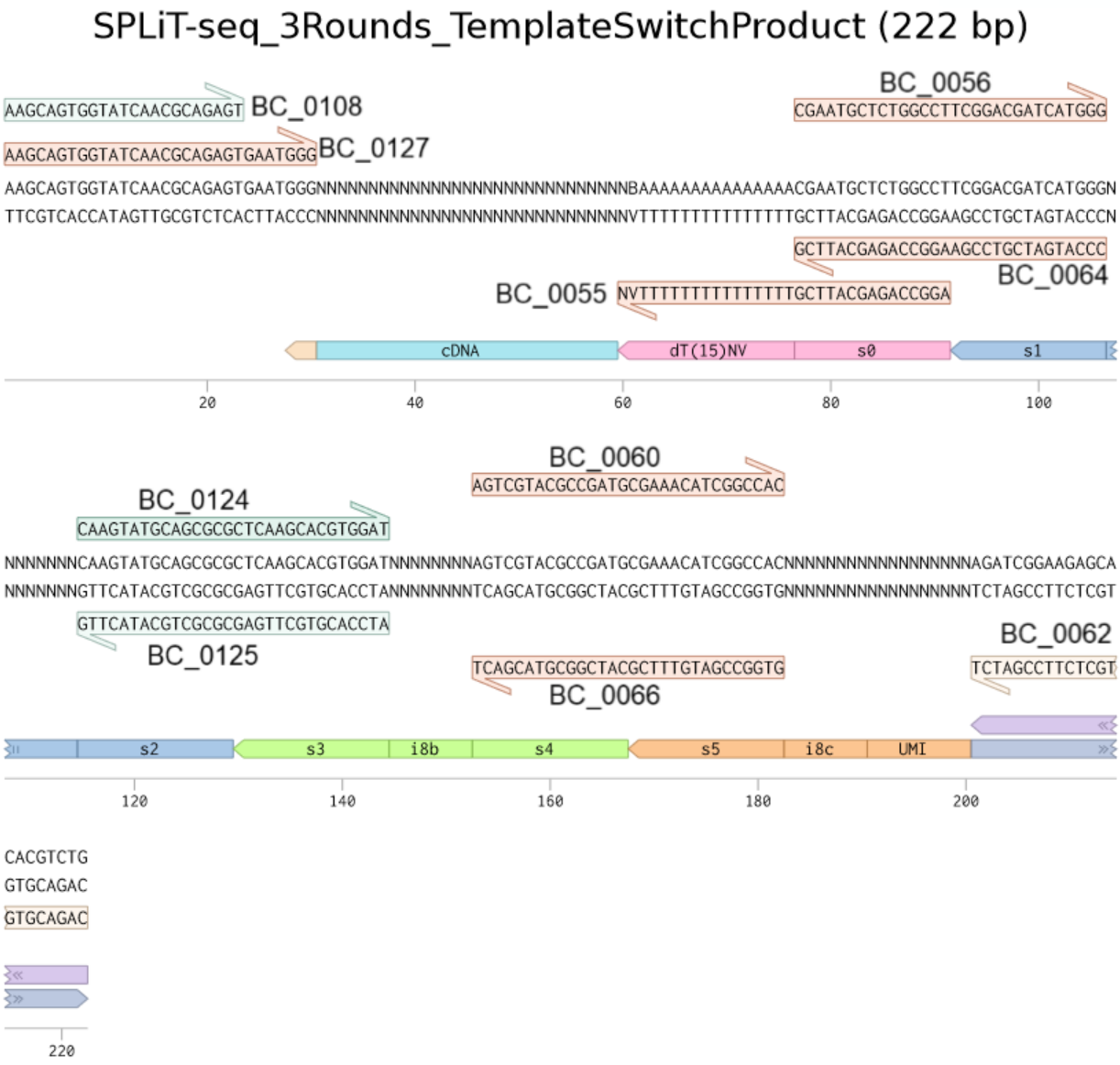


Fig. S24. Sequence map for molecules after template switching in SPLiT-seq protocol.

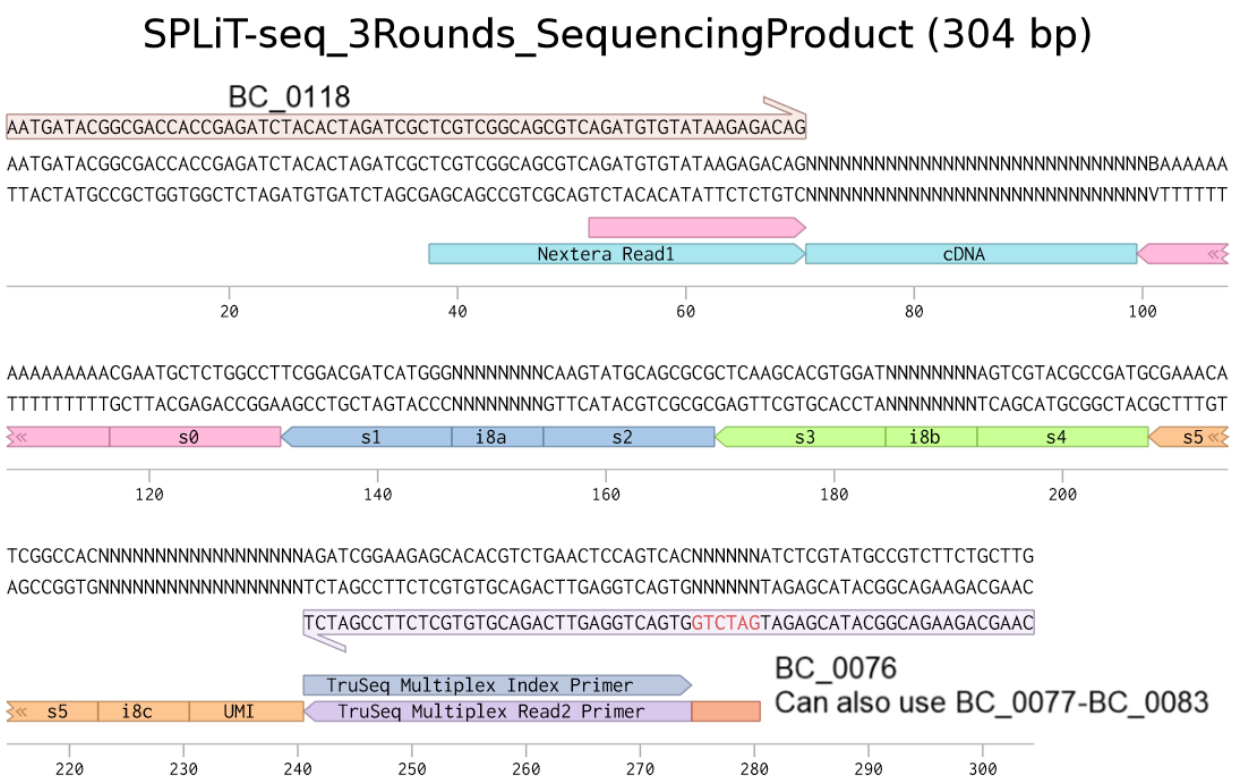


Fig. S25. Sequence map for sequencing-ready molecules after tagmentation in SPLiT-seq protocol.

Supplementary Tables

	Total Cells	Human Cells	Mouse Cells	Mixed Cells	Fraction Human Cells	Fraction Mouse Cells	Fraction Mixed Cells	Mean Species Purity - Human	Mean Species Purity - Mouse	Median UMIs/UBC - Human	Median UMIs/UBC - Mouse	Median Genes/UBC - Human	Median Genes/UBC - Mouse	UMI Duplication
Whole Cells														
Fresh, Sample 1	1758	868	888	2	0.4937	0.5051	0.0011	0.9963	0.9919	9146.5	6808.5	4183	3253.5	66.78%
Fresh, Sample 2	168	80	88	0	0.4762	0.5238	0	0.9958	0.9901	15365	12243	5498	4497	94.54%
Frozen, Sample 1	583	293	289	1	0.5026	0.4957	0.0017	0.9953	0.9921	8363	6702	4046	3231	74.29%
Frozen, Sample 2	94	43	50	1	0.4574	0.5319	0.0106	0.9944	0.9883	15078	10951.5	5540	4319	93.43%

	Total Nuclei	Human Nuclei	Mouse Nuclei	Mixed Nuclei	Fraction Human Nuclei	Fraction Mouse Nuclei	Fraction Mixed Nuclei	Mean Species Purity - Human	Mean Species Purity - Mouse	Median UMIs/UBC - Human	Median UMIs/UBC - Mouse	Median Genes/UBC - Human	Median Genes/UBC - Mouse	UMI Duplication
Nuclei														
Fresh, Sample 1	1488	695	757	36	0.4671	0.5087	0.0242	0.9961	0.9925	9716	4847	4140	2566	66.78%
Fresh, Sample 1, Filtered	471	407	42	22	0.8641	N/A	N/A	0.9955	N/A	8193	N/A	3822	N/A	66.78%
Fresh, Sample 2	144	66	75	3	0.4583	0.5208	0.0208	0.9959	0.9922	15652	8467	5417.5	3607	94.54%
Fresh, Sample 2, Filtered	39	32	6	1	0.8205	N/A	N/A	0.9958	N/A	12113	N/A	4663	N/A	94.54%
Frozen, Sample 1	585	252	319	14	0.4308	0.5453	0.0239	0.9960	0.9930	12564.5	6162	4819	2998	74.29%
Frozen, Sample 1, Filtered	159	140	14	5	0.8805	N/A	N/A	0.9949	N/A	10489.5	N/A	4365.5	N/A	74.29%
Frozen, Sample 2	109	52	54	3	0.4771	0.4954	0.0275	0.9943	0.9911	16815	9883	5411	4020.5	93.43%
Frozen, Sample 2, Filtered	40	35	3	2	0.8750	N/A	N/A	0.9930	N/A	13636	N/A	4982	N/A	93.43%

Note 1: Two different samples were processed for each of the 4 conditions (fresh whole cells, frozen/stored whole cell, fresh nuclei, and frozen/stored nuclei). In all cases, sample 1 consists of more cells with lower sequencing depth and sample 2 consists of less cells with higher sequencing depth.
Note 2: Nuclei samples were filtered to contain less than 1% mitochondrial reads. In some cases, nuclei extraction on cell lines was inefficient, yielding a mixture of whole cells and nuclei. For this reason, statistics for both unfiltered and filtered nuclei are shown.

Table S1. Summary of species-mixing experiments with fresh/frozen whole cells/nuclei. Metrics of species-mixing experiments performed on SPLiT-seq. "Fresh" indicates that cells were harvested and directly processed using the SPLiT-seq workflow. "Frozen" indicates that cells were harvested, fixed and stored for 2 weeks at -80˚C before continuing with the SPLiT-seq workflow. Nuclei samples underwent a computational filtering step where any uniquely barcoded nuclei containing > 1% mitochondrial reads was removed ("Filtered" samples). Due to insufficient nuclei extraction from NIH/3T3 (mouse) cells, metrics for mouse samples in the filtered nuclei datasets have been excluded.

	Total Nuclei	Mean Genes/nucleus	Mean UMIs/nucleus	Median Genes/nucleus	Median UMIs/nucleus	Total Reads	Raw reads/nucleus	UMI Duplication Rate
Small Library	131	2,729.92	11,701.03	2,055	4,943	36,332,629	277,348	94.60%
Large Library	163,069	797.96	1,347.19	677	1,022	2,456,381,771	15,063	58.20%

Table S2. Summary of snRNA-seq on mouse central nervous system.

Cluster Number	Cluster Name	Class	Number Cells	Marker Genes from Literature	Reference
1	Olfactory Bulb Mitral and Tufted - Eomes	Neuron	117	*Tbx21*	(*89*, *90*)
2	Olfactory Bulb Mitral and Tufted - Ms4a15	Neuron	271	*Tbx21*	(*89*, *90*)
3	Olfactory Bulb Mitral and Tufted - Svil	Neuron	89	*Tbx21*	(*89*, *90*)
4	Striatal Medium Spiny Neurons	Neuron	6106	*Drd2*, *Ppp1r1b*	(*91*, *110*)
5	Cortex Layer2/Layer 3 Pyramidal - Satb1	Neuron	69	*Ntf3*	(*111*)
6	Cortex Layer 2/Layer3/Layer 4 Pyramidal - Ntf3	Neuron	1446	*Rasgrf2*	(*11*)
7	Cortex Layer 2/Layer 3 Pyramidal Met	Neuron	1880	*Rasgrf2, Pvrl3, Cux2*	(*11*, *112*)
8	Cortex Layer 4 Pyramidal - Wnt5b	Neuron	255	*Slc17a6, Satb2, Sema3c*	(*11*)
9	Cortex Layer 2/Layer3/Layer 4 Pyramidal - Mef2c	Neuron	8332	*Rasgrf2, Pvrl3, Cux2, Rorb*	(*11*)
10	Cortex Layer 4 Pyramidal - Rorb	Neuron	779	*Thsd7a, Rorb, Cux2, Pvrl3, Rasgrf2*	(*11*, *112*)
11	Cortex Layer 4/Layer 5 Pyramidal	Neuron	2831	*Thsd7a*	(*11*)
12	Cortex Layer 5 Pyramidal - Itbg3	Neuron	198	*Rorb, Thsd7a, Sulf2, Kcnk2,Grik3, Etv1*	(*11*, *112*)
13	Cortex Layer 5 Pyramidal - Fezf2	Neuron	352	*Kcnk2, Grik3, Foxp2, Tle4, Tmem200a, Glra2, Etv1*	(*11*, *112*)
14	Cortex Layer 6a Pyramidal	Neuron	1498	*Grik3*	(*112*)
15	Cortex Layer 5/Layer 6 Pyramidal - Sulf1	Neuron	493	*Sulf2,Grik3, Tle4, Htr1f, Sulf1*	(*11*, *112*)
16	Cortex Layer 5/Layer 6 Pyramidal - Npr3	Neuron	550	*Grik3, Tle4, Rxfp1*	(*11*, *112*)
17	Cortext Layer 6 Pyramidal - Htr1f	Neuron	3479	*Syt6, Grik3, Foxp2, Tle4, Htr1f*	(*11*, *112*)
18	Claustrum Pyramidal	Neuron	140	*Nr4a2*	(*11*)
19	Mesencephalic Tectum Glutamatergic	Neuron	736	*Tfap2d, Slc17a6*	(*93*)
20	Thalamic Gluatamatergic	Neuron	2627	*Lef1, Tcf7l2,Cacna1g,Slc17a6,Wnt3*	(*113*, *114*)

21	Thalamic Interneuron	Neuron	202	*Six3, Gad1, Gad2*	(*115*)
22	Purkinje Early	Neuron	208	*Pcp2, Pde1c*	(*91*)
23	Purkinje Late	Neuron	171	*Pcp2, Slc9a3*	(*91*)
24	Cerebellar Interneuron Progenitor	Neuron	160	*Pax3, Mki67*	(*42*, *91*)
25	Cerebellar Granule Precursor	Neuron	4364	*Gli2*	(*116*)
26	Cerebellar Interneuron - Stellate and Basket	Neuron	459	*Rora*	(*42*, *91*)
27	Cerebellar Interneuron - Golgi, Stellate and Basket	Neuron	420	*Tfap2b*	(*42*, *91*)
28	Cerebellar Granule	Neuron	10996	*Gabra6*	(*117*)
29	Cerebellar Interneuron Precursor	Neuron	851	*Pax2*	(*42*, *91*)
30	Medulla Glycinergic Interneuron - Rxfp2	Neuron	329	*Stac, Slc6a5, Glra1*	(*118*)
31	Medulla Interneuron - Rxfp2	Neuron	282	*Stac*	(*118*)
32	Nigral Dopaminergic	Neuron	47	*Slc6a3*	(*119*)
33	Hippocampal Pyramidal - Cr2	Neuron	232	*Cr2*	(*120*)
34	Subiculum Pyramidal	Neuron	78	*Ntm, Rxfp1, Nr4a2*	(*121*)
35	Hippocampal Pyramidal - Crym	Neuron	1260	*Crym*	(*122*)
36	Hippocampus Granule Progenitor Mki67	Neuron	657	*Prox1, Mki67*	(*32*)
37	Hippocampal Pyramidal Precursor	Neuron	315	*Nrp1, Zbtb20*	(*123*, *124*)
38	Hippocampal Pyramidal - Grik4	Neuron	625	*Grik4, Slc17a7*	(*125*)
39	Hippocampal Granule Precursor - Nrp2	Neuron	515	*Prox1, Nrp2*	(*32*)
40	Hippocampal Granule/Pyramidal CA3	Neuron	772	*Prox1, Slc17a7*	(*32*)
41	Hippocampal Pyramidal - Npy2r	Neuron	117	*Npy2r, Slc17a7*	(*126*)
42	Spinal Cord Glutamatergic - Hmga2	Neuron	231	*Slc17a6, Slc17a8*	(*127*)
43	Spinal Cord Glutamatergic Gna14	Neuron	95	*Slc17a8, Gna14*	(*127*)

44	Migrating Interneuron - Lhx6	Neuron	3907	*Dlx family, Lhx6*	(*128*)
45	Migrating Interneuron - Trdn	Neuron	246	*Dlx family, Trdn*	(*128*)
46	Migrating Interneuron - Cpa6	Neuron	2166	*Dlx family, Cpa6*	(*128*)
47	Migrating Interneuron - Foxp2	Neuron	701	*Dlx family, Foxp2*	(*128*)
48	Migrating Interneuron - Pbx3	Neuron	1835	*Dlx family, Pbx3*	(*128*)
49	Migrating Interneuron - Lgr6	Neuron	484	*Dlx family, Lgr6*	(*128*)
50	Migrating Interneuron - Adarb2	Neuron	47	*Dlx family, Adarb*	(*128*)
51	Subventricular Zone Stem Cell	Neuron	182	*Gfap, Vim, Nes, Dlx family*	(*129*)
52	Cajal-Retzius	Neurons	133	*Trp73, Reln*	(*130*)
53	Unresolved	Neuron	36469		
54	Unresolved	Neuron	90		
55	Oligodendroctye Myelinating 2	Oligo -dendrocyte	721	*Opalin*	(*12*)
56	Oligodendroctye Myelinating 1	Oligo -dendrocyte	1781	*Opalin*	(*12*)
57	Oligodendrocyte Mature	Oligo -dendrocyte	191	*Hapln2*	(*12*)
58	Oligodendroctye Newly Formed 1	Oligo -dendrocyte	467	*Tmem2*	(*12*)
59	Committed Oligodendroctye Precursor Cells 1	Oligo -dendrocyte	811	*Gpr17*	(*12*)
60	Committed Oligodendroctye Precursor Cells 2	Oligo -dendrocyte	323	*Bcas1*	(*12*)
61	Oligodendroctye Precursor Cells	OPC	5793	*Pdgfra*	(*12*)
62	Perivivascular Macrophage	Immune	63	*Dab2*	(*131*)
63	Microglia	Immune	558	*Tgfbr1*	(*132*)
64	Endothelia	Vasc -ulature	561	*Flt1, Kdr*	(*10*)
65	Smooth Muscle Cells	Vasc -ulature	98	*Abcc9, Pdgfrb*	(*96*)
66	Vascular and Leptomeningeal Cells 2	VLMC	1223	*Slc6a13*	(*12*)
67	Vascular and Leptomeningeal Cells 1	VLMC	251	*Slc47a1, Slc47a2*	(*12*)
68	Astrocyte - Slc7a10	Astrocyte	3569	*Slc7a10*	(*133*)
69	Astrocyte - Prdm16	Astrocyte	8103	*Prdm16*	(*134*)
70	Astrocyte - Gfap	Astrocyte	282	*Gfap*	(*135*)
71	Bergman Glia	Astrocyte	1527	*Gria1*	(*136*)
72	Ependyma	Ependyma	518	*Dnah1/2/5/9/10/11*	(*11*)

73	Olfactory Ensheathing Cells (OEC)	Schwann Cells	256	*Lama4, Col5a2, Runx1*	(*137*, *138*)

Table S3. Table of assigned cell types and marker genes from literature

Table S4. Top 50 differentially expressed genes in each cluster from the joint brain and spinal cord clustering. Differential expression is calculated as $\log2(TPM_{CLUSTER}+1)/ \log2(TPM_{\sim CLUSTER}+1)$, where $TPM_{\sim CLUSTER}$ is the average TPM for all the cells not in the cluster of interest. We only include genes expressed in at least 20% of the transcriptomes in a cluster.

Table S5. Average expression for each cluster from the joint brain and spinal cord clustering. All values are listed as TPM+1.

Table S6. Genes used to generate P4 sagittal composite ISH maps. Top ten differentially expressed genes from each cluster that were also available in the Allen ISH database for a postnatal day 4 mouse were used.

Table S7. Genes used to generate P14 sagittal composite ISH maps. Top ten differentially expressed genes from each cluster that were also available in the Allen ISH database for a postnatal day 4 mouse were used.

Cluster Number	Cluster Name	Number Cells	Marker Genes from Literature	Ref
1	Ependymal	139	*Dnah1/2/5/9/10/11*	(*11*)
2	Unassigned	58		
3	Unassigned	44		
4	Astroctye - Unassigned	116		
5	Astrocyte - Gfap	394	*Aldh1l1*	(*139*)
6	Astrocyte - Slc7a10	1230	*Aldh1l1*	(*139*)
7	Astro - Svep1	986	*Aldh1l1*	(*139*)
8	Endothelial	95	*Aldh1l1*	(*139*)
9	VLMC	444	*Col1a2*	(*12*)
10	Microglia	91	*Tgfbr1*	(*132*)
11	Oligodendrocyte Mature	1436	*Mog*	(*11*)
12	Oligodendroctye Myelinating	489	*Tmem2*	(*11*)
13	OPC	1213	*Pdgfra*	(*11*)
14	Committed OPC	835	*Bcas1*	(*11*)
15	Cerebrospinal Fluid-Contacting Neurons (CSF-cNs)	51	*Pkd2l1, Pkd1l2*	(*44*)
16	Alpha motor neurons	100	*Chat, Esrrg-*	(*45, 46*)
17	Gamma motor neurons	77	*Chat, Esrrg, Esrrb, Htr1d*	(*45, 46*)
18	Inhibitory 1	46	*Gad1, Gad2*	(*140*)
19	Inhibitory 2	365	*Gad1, Gad2*	(*140*)
20	Inhibitory 3	59	*Gad1, Gad2*	(*140*)
21	Inhibitory 4	361	*Gad1, Gad2*	(*140*)
22	Inhibitory 5	397	*Gad1, Gad2*	(*140*)
23	Inhibitory 6	289	*Gad1, Gad2*	(*140*)
24	Inhibitory 7	220	*Gad1, Gad2*	(*140*)
25	Inhibitory 8	81	*Gad1, Gad2*	(*140*)
26	Inhibitory 9	321	*Gad1, Gad2*	(*140*).
27	Inhibitory 10	40	*Gad1, Gad2*	(*140*)
28	Excitatory 1	54	*Slc17a6*	(*141*)
29	Excitatory 2	450	*Slc17a6*	(*141*)
30	Excitatory 3	185	*Slc17a6*	(*141*)
31	Excitatory 4	365	*Slc17a6*	(*141*)
32	Unresolved	7634		
33	Unresolved	65		
34	Excitatory 5	53	*Slc17a6*	(*141*)
35	Excitatory 6	243	*Slc17a6*	(*141*)
36	Excitatory 7	41	*Slc17a6*	(*141*)
37	Excitatory 8	189	*Slc17a6*	(*141*)
38	Excitatory 9	389	*Slc17a6*	(*141*)
39	Excitatory 10	385	*Slc17a6*	(*141*)
40	Excitatory 11	85	*Slc17a6*	(*141*)
41	Excitatory 12	612	*Slc17a6*	(*141*)
42	Excitatory 13	166	*Slc17a6*	(*141*)
43	Excitatory 14	597	*Slc17a6*	(*141*)
44	Excitatory 15	144	*Slc17a6*	(*141*)

Table S8. Table of assigned spinal cord cell types and marker genes from literature

Table S9. Top 50 differentially expressed genes in each cluster from the spinal cord clustering. Differential expression is calculated as $\log2(TPM_{CLUSTER}+1)/ \log2(TPM_{\sim CLUSTER}+1)$, where $TPM_{\sim CLUSTER}$ is the average TPM for all the cells not in the cluster of interest. We only include genes expressed in at least 20% of the transcriptomes in a cluster.

Table S10. Average expression for each cluster from the spinal cord clustering. All values are listed as TPM+1.

Items	Supplier	Item Code	Cost Per Experiment (USD)
Maxima H Minus	ThermoFisher	EP0753	386.28
RNase Inhibitor	Enzymatics, Ambion	Y9240L, AM2696	69.20
T4 DNA Ligase	New England Biolabs	M0202L	307.20
Kapa Pure Beads	Kapa Biosystems	KK8002	12.00
Dynabeads MyOne C1	ThermoFisher	65002	1.70
Nextera XT DNA Preparation Kit	Illumina	FC-131-1096	28.90
Kapa Hotstart HiFi ReadyMix	Kapa Biosystems	KK2602	22.26
Proteinase K	ThermoFisher	EO0491	3.44
dNTPs	ThermoFisher	R0192	5.03
Oligonucleotides	Integrated DNA Technologies, Exiqon	N/A	37.05
Total			873.07

Table S11. Itemized cost breakdown of SPLiT-seq

SPLiT-seq Protocol

Projected Experimental Time: 2 Days

Recommended time on day 1 to start: morning

Addition of Rnase inhibitors to buffers:

When any buffer has "+RI" next to it, this indicates that enzymatics RNase inhibitor should be added to a final concentration of 0.1 U/uL.

Centrifugation Steps

All centrifugation steps should be performed with a swinging bucket rotor. Using a fixed angle centrifuge may lead to more cell loss. Depending on the tissue type, centrifugation speeds may need to be changed to optimize cell retention (e.g. smaller cells = higher speeds).

DNA Barcoding Plate Generation

What you need:

- Three 96 well plates from IDT - Reverse Transcription Barcode Primers, Ligation Round 1,and Ligation Round 2 Stock DNA Oligo plates (100 uM)
- Two linker oligos - BC_0215, BC_0060 (*Note: these are assumed to be in stock concentration of 1mM, be sure to correct for volume if only have 100 uM stocks)*
- Six 96 well PCR plates (3 stock plates that will last at least 10 experiments, and 3 plates for 1st experiment)

Note: This will generate 100 uL of DNA barcodes for each well. Each SPLiT-seq experiment requires only 4 uL/well of the reverse transcription primer solution which will last for 25 experiments. Each SPLiT-seq experiment requires only 10 uL/well of the barcode/linker solutions, so these plates will last a total of 10 experiments.

Round 1 reverse transcription barcoded primers (final concentrations of 12.5 uM random hexamer and 12.5 uM 15dT primers in each of 48 wells)

1. Using multichannel pipette, add 12.5 uL of rows A-D in the IDT Reverse Transcription Barcode Primers to rows A-D of the BC Stock 96 well PCR plate.
2. Using multichannel pipette, add 12.5 uL of rows E-H in the IDT Reverse Transcription Barcode Primers to rows A-D of the BC stock 96 well PCR plate (mixing polydT with random hexamer primer here)
3. Add 75ul of water to rows A-D of the BC stock 96 well PCR plate.

Round 2 ligation round (Final concentrations of 12uM barcodes, 11uM linker-BC_0215)

1. Using multichannel pipette, add 12uL of IDT Round 2 Barcodes to R1 Stock 96 well PCR plate

2. Add 138.6ul of BC_0215(1mM) to 10.9494mL water in a basin (BC_0215_dil)
3. Using multichannel pipette, add 88uL BC_0215_dil to each well of R2 Stock 96 well PCR plate

Ligation Round 3 (Final concentrations of 14uM barcodes, 13uM linker-BC_0060)

1. Using multichannel pipette, add 14uL of Round 3 Barcodes to R3 Stock 96 well PCR plate
2. Add 163.8ul of BC_0060(1mM) to 10.6722mL water in a basin (BC_0060_dil)
3. Using multichannel pipette, add 86uL BC_0060 to each well R3 Stock 96 well PCR plate

For each ligation plate (R2 and R3, not including reverse transcription barcodes), anneal the barcode and linker oligos with the following thermocycling protocol:

1. Heat to 95C for 2 minutes
2. Ramp down to 20C for at a rate of -0.1C/s
3. 4C

Aliquot out 10 uL of each barcode/linker stock plate into 3 new 96 well PCR plates. These are the plates that should be used for DNA barcoding in the split-pool ligation steps in the protocol.

Nuclei Extraction (Optional):

1. Prepare the following items:
    - Keep dounce at 4C until use
    - 15ml of 1xPBS + 37.5 Superase-in + 19ul Enzymatics Rnase inhibitor. (kept on ice)
    - Precool centrifuge to 4C
2. Make **NIM1 buffer**:

Reagent	Stock Concentration	Final Concentration	Volume (uL)
Sucrose	1.5 M	250mM	2,500
KCl	1 M	25mM	375
MgCl2	1 M	5mM	75
Tris buffer, pH 8	1 M	10mM	150
Water	NA	NA	11,900
Final Volume			15,000

3. Make the **homogenization buffer:**

Reagent	Stock Concentration	Final Concentration	Volume (uL)
NIM1 Buffer	1.5 M		4,845
1 mM DTT	1 mM	1uM	5
Enzymatics RNase-In (40U/ul)	40 U/uL	0.4U/ul	50
Superase-In (20U/UL)	20 U/uL	0.2U/ul	50
10% Triton X-100	10%	NA	50
Final Volume			5,000

4. Dounce
    - Add tissue/cells sample to dounce. If cells, resuspend in 700ul of homogenization buffer.
    - Add homogenization buffer to ~700ul
    - Perform 5 strokes of loose pestle
    - Perform 10 - 15 of tight pestle
    - Add homogenization buffer up to 1ml
    - Check cell lysis with 5ul trypan blue and 5ul cells on haemocytometer to see if nuclei have been released
5. Filter homogenates with 40um strainer into 5ml eppendorf tubes (or 15mL falcon). Tilting the filter 45° while straining over the tube ensures that the lysate passes through as intended.

    Note: This straining process is different from every other one below.
6. Spin for 4min at 600g (4C) and remove supernatant (can leave about 20uL to avoid aspirating pellet)
7. Resuspend in 1ml of 1x PBS + RI
8. Add 10ul of BSA
9. Centrifuge at 600g for 4min.
10. Resuspend in 200ul 1x PBS + RI.
11. Take 50ul of the resuspended cells from step 4 and add 150ul of 1xPBS + RI. Count sample on hemocytometer and/or flow-cytometer.
    - The volume of resuspended cells from the step 4 can be changed based on the considerations of the user.
12. Pass cells through a 40um strainer into a fresh 15mL Falcon tube and place on ice.
    - See note on step 4 of Fixation and Permeabilization.
13. Resuspend the desired number of nuclei (typically 2M) in 1mL 1x PBS + RI and proceed with step 5 in the following *Fixation and Permeabilization* protocol.

Fixation and Permeabilization

1. Prepare the following buffers (calculated for two experiments):
    - A 1.33% formalin (360 uL of 37% formaldehyde solution (Sigma)+ 9.66 ml PBS) solution and store at 4C.
    - 6 mL of 1X PBS+RI (15 uL of SUPERase In and 7.5 uL of Enzymatics RNase inhibitor)
    - 2 mL of 0.5X PBS+RI (5 uL of SUPERase in and 2.5 of Enzymatics RNase inhibitor)
    - 500uL of 5% Triton X-100 + RI (2 uL of SUPERase In)
    - 1100uL of 100mM Tris pH 8.0 + 4 uL SUPERase In
    - Set the centrifuge to 4C
2. Pellet cells by centrifuging at 500g for 3 mins at 4C. (Some cells may require faster centrifugation.)
3. Resuspend cells in 1mL of cold PBS+RI. Keep cells on ice between these steps.
4. Pass cells through a 40um strainer into a fresh 15mL Falcon tube and place on ice.

    Note: The cell resuspension is not likely to passively go through the strainer, which can cause cell loss. Instead, with a 1ml pipette filled with the resuspension, press the end of the tip directly onto the strainer and actively push the liquid through. The motion should take ~1 second.

5. Add 3 mL of cold 1.33% formaldehyde (final concentration of 1% formaldehyde). Fix cells on ice for 10 mins.
6. Add 160uL of 5% Triton-X100+RI to fixed cells and mix by gently pipetting up and down 5x with a 1mL pipette. Permeabilize cells for 3 mins on ice.
7. Centrifuge cells at 500g for 3 mins at 4C.
8. Aspirate carefully and resuspend cells in 500 uL of cold PBS+RI.
9. Add 500uL of cold 100 mM Tris-HCl, pH 8.0.
10. Add 20 uL of 5% Triton X-100.
11. Centrifuge cells at 500g for 3 mins at 4C.
12. Aspirate and resuspend cells in 300 ul of cold **0.5x** PBS+RI.
13. Run cells through a 40uM strainer into a new 1.7mL tube.
    - See note on step 4 of Fixation and Permeabilization.
14. Count cells using a hemacytometer or a flow-cytometer and dilute the cell suspension to 1,000,000 cells/mL. While counting cells, keep cell suspension on ice.

    Note: This step will dictate how many cells enter the split-pool rounds. It will be possible to sequence only a subset of the cells that enter the split-pool rounds (can be done during sublibrary generation at lysis step). The total number of barcode combinations you will be using should be calculated to determine the maximum number of cells you can sequence with minimal barcode collisions. As a rule of thumb, the number of cells you process should not exceed more than 5% of total barcode combinations. We usually have a dilution between 500k to 1M cells/mL here (equates to 4-8k cells going into each well for reverse transcription barcoding rounds).

Reverse Transcription

1. Aliquot out 4 uL of the RT barcodes stock plate into the top 4 rows (48 wells) of a new 96 well plate. Cover the this plate with an adhesive plate seal until ready for use.
2. Create the following reverse transcription (RT) mix on ice:

Reagent	Stock Concentration	Desired Concentration	Per Reaction	Volume in Mix (48 wells + 10%)
5X RT Buffer	5x	1x	4	211.2
Enzymatics Rnase Inhibitor	40u/uL	0.25u/uL	0.125	6.6
Superase In Rnase Inhibitor	20U/uL	0.25U/uL	0.25	13.2
dNTPs	10mM (per base)	500uM	1	52.8
Maxima H Minus Reverse Transcriptase	200u/uL	20u/ul	2	105.6
H2O	NA	NA	0.625	33
Total Volume			**8**	**422.4**

3. Add 8uL of the RT mix to each of the top 48 wells. Each well should now contain a volume of 12uL.
4. Add 8uL of cells in 0.5x PBS+RI to each of the top 48 wells. Each well should now contain a volume of 20uL.
5. Add the plate into a thermocycler with the following protocol
    a. 50 C for 10 minutes
    b. Cycle 3 times:
        i. 8C for 12s
        ii. 15C for 45s
        iii. 20C for 45s
        iv. 30C for 30s
        v. 42C for 2 min
        vi. 50C for 3 min
    c. 50C for **5 min**
    d. 4C forever
6. Place the RT plate on ice.
7. Prepare 2 mL of 1x NEB buffer 3.1 with 20uL of Enzymatics RNase Inhibitor.
8. Transfer each RT reaction to a 15mL falcon tube (also on ice).
9. Add 9.6uL of 10% Triton-X100 to get a final concentration of 0.1%.
10. Centrifuge pooled RT reaction for 3 min at 500G.

11. Aspirate supernatant and resuspend into 2 mL of 1x NEB buffer 3.1 + 20uL Enzymatics RNase Inhibitor.

Ligation Barcoding

Make the following ligation master mix on ice:

Note: Final concentration takes added volume of DNA barcodes into account. Concentrations of this mix is not the final concentration at time of barcoding

Reagent	Stock Concentration	Final Concentration	Volume (uL)
Water	NA	NA	1337.5
T4 Ligase Buffer 10x	10X	1X	500
Enzymatics Rnase Inhibitor	40 U/uL	0.32 U/uL	40
Superase In	20 U/uL	0.05 U/uL	12.5
BSA	20 mg/mL	0.2 mg/mL	50
T4 DNA Ligase	400 U/uL	8 U/uL	100
Total Volume			**2040**

1. Add the 2mL of cells in NEB buffer 3.1 into the ligation mix. The mix should now have a volume of 4.04 mL
2. Add the mix into a basin
3. Using a multichannel pipet, add 40 uL of ligation mix (with cells) into each well of the round 1 DNA barcode plate.
4. Cover the round 1 DNA barcode plate with an adhesive plate seal and incubate for **30 minutes at 37C** with gentle rotation (50 rpm).
5. Make the round 1 blocking solution and add it to a new basin

Reagent	Stock Concentration	Final Concentration	Volume (uL)
BC_0216	100 uM	26.4 uM	316.8
10x Ligase Buffer	10X	2.5X	300
Water	NA	NA	583.2
Final Volume			**1200 uL**

6. Remove the round 1 DNA barcoding plate from the incubator and remove the cover.
7. Using a multichannel pipet, add 10 uL of the round 1 blocking solution to each of the 96 wells in the round 1 DNA barcoding plate.
8. Cover the round 1 DNA barcode plate with an adhesive plate seal and incubate for **30 minutes at 37C** with gentle rotation (50 rpm).
9. Remove round 1 DNA barcoding plate from the incubator, remove cover, and pool all cells into a new basin.
10. Pass all the cells from this basin through a 40 um strainer into another basin.
    - See note on step 4 of Fixation and Permeabilization.
11. Add 100 uL of T4 DNA ligase to the basin and mix by pipetting ~20 times.
12. Using a multichannel pipette, add 50 uL of cell/ligase solution into each well of the round 2 DNA barcode plate.
13. Cover the round 2 DNA barcode plate with an adhesive plate seal and incubate for **30 minutes at 37C** with gentle rotation (50 rpm).
14. Make the round 2 blocking solution and add it to a new basin

Reagent	Stock Concentration	Final Concentration	Volume (uL)
BC_0066	100 uM	11.5 uM	369
EDTA	0.5 M	125 mM	800
Water	NA	NA	2031
Final Volume			**3200 uL**

15. Remove the round 2 DNA barcoding plate from the incubator and remove the cover.
16. Using a multichannel pipet, add 20 uL of the round 2 blocking and termination solution to each of the 96 wells in the round 2 DNA barcoding plate.
17. Pool all cells into a new basin. (no incubation for the final blocking step)
18. Pass all the cells from this basin through a 40 um strainer into a 15 mL falcon tube.
    - See the note for step 4.
19. Count cells on a flow cytometer. Make sure cells are well mixed before aliquoting sample for counting.

Lysis

1. Make the 2X lysis buffer:

Reagent	Stock Concentration	Final Concentration (2X)	Volume (mL)
Tris, pH 8.0	1 M	20 mM	0.5
NaCl	5 M	400 mM	2
EDTA, pH 8.0	0.5 M	100 mM	5
SDS	10%	4.4 %	11
Water	NA	NA	6.5
Final Volume			**25**

2. If white precipitate appears, warm at 37C until precipitate is back in solution (roughly 10-15 min).
3. Make the following wash buffer:

Reagent	Volume (uL)
1X PBS	4000
10 % Triton X-100	40
Superase In Rnase Inhibitor	10
Final Volume	**4050**

4. Add 70ul of 10% triton to the cells. (~0.1% final conc.)
5. Centrifuge for 5 min at 1000G in 15ml tube.
   Note: The pellet for the steps below will be very small and it may not be visible.
6. Aspirate supernatant, leave ~30ul to avoid removing pellet.
   a. If possible, remove as much supernatant as possible with 20uL pipet.
7. Resuspend with 4 mL of wash buffer.
8. Centrifuge for 5 min at 1000G.
9. Aspirate supernatant and resuspend in 50ul 1x PBS + RI.
10. Dilute 5ul into 195uL of 1x PBS and count via flow cytometry.
    - Or take 5ul into 5ul of 1x PBS and count on hemocytometer (it can be hard to distinguish debris from cells).

11. Determine how many sublibraries you would like to generate (# sublibraries= # tubes needed), and how many cells you would like to have for each of these sublibraries.
12. Aliquot the desired number of cells for each sublibrary into new 1.7mL tubes. Add 1x PBS to each tube to a final volume of 50uL.
13. Add 50uL of 2x Lysis buffer to each tube.
14. Add 10uL of Proteinase K (20mg/mL) to each lysate.
15. Incubate at 55C for 2 hrs with shaking at 200rpm.
16. Stopping point: Freeze lysate(s) at -80C.

Prepare buffers

First make the following stock solutions:

100mM PMSF (resuspended in isopropanol)

2x B&W	
Reagents	**Volume**
1M Tris-HCl pH 8.0	500uL
5M NaCl	20ml
EDTA, 0.5M	100ul
Nuclease Free Water	29.4ml
Total	**50mL**

1x B&W-T	
Reagents	**Volume**
1M Tris-HCl pH 8.0	100uL
5M NaCl	4ml
EDTA, 0.5M	20ul
Tween 20 10%	100ul
Nuclease Free Water	15.78ml
Total	**20mL**

Then make the following smaller aliquots (with added RNase inhibitor):

1x B&W-T + RI:

	Volume per Number of Samples (uL)							
Reagent	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1xB&W-T	3600.0	4200.0	4800.0	5400.0	6000.0	6600.0	7200.0	7800.0
SUPERase In	5.0	5.8	6.7	7.5	8.3	9.2	10.0	10.8
Final Volume	**3605.0**	**4205.8**	**4806.7**	**5407.5**	**6008.3**	**6609.2**	**7210.0**	**7810.8**

2x B&W + RI:

	Volume per Number of Samples (uL)							
Reagent	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
2xB&W	110.0	220.0	330.0	440.0	550.0	660.0	770.0	880.0
SUPERase In	2.0	4.0	6.0	8.0	10.0	12.0	14.0	16.0
Final Volume	**112.0**	**224.0**	**336.0**	**448.0**	**560.0**	**672.0**	**784.0**	**896.0**

Tris-T + RI:

	Volume per Number of Samples (uL)							
Reagent	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
10mM Tris-HCl (pH 8.0)	600.0	1200.0	1800.0	2400.0	3000.0	3600.0	4200.0	4800.0
Tween-20 (10%)	6.0	12.0	18.0	24.0	30.0	36.0	42.0	48.0
SUPERase In	1.5	3.0	4.5	6.0	7.5	9.0	10.5	12.0
Final Volume	**607.5**	**1215.0**	**1822.5**	**2430.0**	**3037.5**	**3645.0**	**4252.5**	**4860.0**

Purification of cDNA

Note: We performed agitation steps on a vortexer with a foam 1.7mL tube holder on a low setting (2/10).

Wash MyOne C1 Dynabeads

1. For each lysate to be processed, add 44uL of MyOne C1 Dynabeads to a 1.5 mL tube (eg, 1 lysate=44uL, 2 lysates = 88uL, 3 lysates = 132ul etc)
2. Add 800uL of 1xB&W-T buffer
3. Place sample against a magnetic rack and wait until liquid becomes clear (1-2 min).
4. Remove supernatant and resuspend beads in 800uL of 1xB&W-T buffer.
5. Repeat steps 3-4 two more times for a total of 3 washes.
6. Place sample against a magnetic rack and wait until liquid becomes clear.
7. Resuspend beads in 100uL (per sample) 2xB&W buffer + RI.

Sample Binding to Streptavidin:

1. Add 5uL of 100uM PMSF (resuspended in isopropanol) to each sample and leave at room temperature for 10 min.
2. Add 100ul of resupended C1 beads to each tube.
3. To bind cDNA to C1 beads, agitate at room temperature for 60 min.
4. Place sample against a magnetic rack and wait until liquid becomes clear (1-2 min).
5. Remove supernatant and resuspend beads in 250uL of 1xB&W-T +RI
6. Agitate beads for 5 min at room temperature.
7. Repeat steps 5 and 6.
8. Remove supernatant and resuspend beads in 250 uL of 10mM Tris-T + RI
9. Agitate beads for 5 min at room temperature.
10. Leave beads in final wash solution on ice.

Template Switch

Prepare the following mix depending on the number of samples:

	Volume per Number of Samples (uL)							
Reagent	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Water	88.0	176.0	264.0	352.0	440.0	528.0	616.0	704.0
Maxima RT Buffer	44.0	88.0	132.0	176.0	220.0	264.0	308.0	352.0
Ficoll PM-400 (20%)	44.0	88.0	132.0	176.0	220.0	264.0	308.0	352.0
10mM dNTPs (each, total is 40mM)	22.0	44.0	66.0	88.0	110.0	132.0	154.0	176.0

RNase Inhibitor	5.5	11.0	16.5	22.0	27.5	33.0	38.5	44.0
TSO (BC_0127)	5.5	11.0	16.5	22.0	27.5	33.0	38.5	44.0
Maxima RT RnaseH Minus Enzyme	11.0	22.0	33.0	44.0	55.0	66.0	77.0	88.0
Total	**220.0**	**440.0**	**660.0**	**880.0**	**1100.0**	**1320.0**	**1540.0**	**1760.0**

1. Place sample against a magnetic rack and wait until liquid becomes clear.
2. With sample still on magnetic rack, remove supernatant and wash with 250uL of water (do not resuspend beads this time).
3. Resuspend sample in 200ul of Template Switch Mix.
4. Incubate at room temp for 30 min with agitation or rolling.
5. Incubate at 42C for 90 min with agitation or rolling (we shook in incubator at 100 rpm).
6. Potential Stopping Point. If stopping perform the following (otherwise skip to next section):
    a. Place sample against a magnetic rack and wait until liquid becomes clear.
    b. Resuspend in 250uL Tris-T.

cDNA Amplification

Prepare the following PCR mix depending on the number of samples:

	Volume per Number of Samples (uL)							
Reagent	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Kapa Hifi 2x Master Mix	121.00	242.00	363.00	484.00	605.00	726.00	847.00	968.00
BC_0108 (10uM)	9.68	19.36	29.04	38.72	48.40	58.08	67.76	77.44
BC_0062 (10uM)	9.68	19.36	29.04	38.72	48.40	58.08	67.76	77.44
Water	101.64	203.28	304.92	406.56	508.20	609.84	711.48	813.12
Total	**242.0**	**484.0**	**726.0**	**968.0**	**1210.0**	**1452.0**	**1694.0**	**1936.0**

1. Place sample against a magnetic rack and wait until liquid becomes clear.
2. With sample against magnet wash with 250uL nuclease-free water (do not resuspend).
3. Resuspend sample with 220uL PCR mix and split equally into 4 different PCR tubes.
4. Run the following thermocycling program:

    a. 95C 3 min
    b. 98C 20s
    c. 65C 45s
    d. 72C 3min
    e. Repeat (b-d) 4x (5 total cycles)
    f. 4C hold.
5. Combine all 4 reactions into a single 1.7mL tube. Make sure to resuspend any beads that may be stuck to the bottom or sides of the PCR tubes before combining reactions.
6. Place sample against a magnetic rack and wait until liquid becomes clear.
7. Transfer 200uL of supernatant to 4 optical grade qPCR tubes (50uL in each tube).
8. Add 2.5uL of 20x evagreen to each qPCR tube.
9. Run the following qPCR program (make sure to remove samples, once signal starts to leave exponential phase to prevent overamplification).
    a. 95C 3 min
    b. 98C 20s
    c. 67C 20s
    d. 72C 3min
    e. Repeat (b-d) until signal plateaus out of exponential amplification
    f. 72C 5 min
    g. 4C hold
10. Optional: Run an agarose gel or bioanalyze resulting qPCR. There will likely be a combination of cDNA and dimer present.

SPRI size selection (0.8x)

1. Combine qPCR reactions into a single tube.
2. Take out 180 uL of the pooled qPCR reaction and place in new 1.7 mL tube
3. Add 144uL of Kapa Pure Beads to tube and vortex briefly to mix. Wait 5 min to bind DNA.
4. Place tube against magnetic rack and wait until liquid becomes clear.
5. Remove the supernatant.
6. With tubes still on magnetic rack, wash with 750uL 85% ethanol. Do not resuspend beads.
7. Repeat step 6.
8. Remove ethanol and air dry bead (~5min). To not let beads overdry and crack.
9. Resuspend beads from each tube in 20uL of water. Once beads are fully resuspended in the water, incubate the tube at 37C for 10 min.
10. Bind tubes against magnetic rack and wait until liquid becomes clear.
11. Transfer 18.5uL of elutant into a new optical grade PCR tube.
12. Run a bioanalyzer trace on 10 uL of the elutant
13. If no dimer is present after size selection, jump directly to “Tagmentation and Illumina Amplicon Generation” section. If dimer is still present, proceed to step 14 to perform a second amplification and size selection step. This may be necessary for cells with low RNA content, but should not be necessary for cells with high RNA content (eg, HeLa-S3, NIH/3T3, etc.).

Tagmentation and Illumina Amplicon Generation

1. Quibit amplified cDNA and dilute to 0.12ng/uL.
2. Preheat a thermocycler to 55 degrees.
3. For each sample, combine 600 pg of purified cDNA with H2O in a total volume of 5 ul.
4. To each tube, add 10 ul of Nextera TD buffer and 5 ul of Amplicon Tagment enzyme (the total volume of the reaction is now 20 ul). Mix by pipetting ~5 times. Spin down.
5. Incubate at 55 C for 5 minutes.
6. Add 5 ul of Neutralization Buffer. Mix by pipetting ~5 times. Spin down. Bubbles are normal.
7. Incubate at room temperature for 5 minutes.
8. Add to each PCR tube in the following order:
    1. 15 ul of Nextera PCR mix
    2. 8 ul H2O
    3. 1 ul of 10 uM (N7 indexed primer, one of BC_0076-BC_0083)
    4. 1 ul of 10 uM Nextera (BC_0118) N501 oligo
9. Run the following thermocycling program:
    1. 95 C 30 sec
    2. 12 cycles of:
        1. 95 C 10 seconds
        2. 55 C 30 seconds
        3. 72 C 30 seconds
    3. Then: 72 C 5 minutes 4 C forever
10. Transfer 40ul out of the 50uL reaction to a 1.7mL tube.
11. Add 28uL of Kapa Pure beads to do a 0.7x cleanup. Elute in 20ul.
12. Bioanalyze resulting sample and quibit before sequencing. See lane 1 on figure 1 for expected size distribution.

Illumina Sequencing

1. Use a paired-end sequencing run with a 150 bp kit.
2. Set read1 to 66 nt (transcript sequence)
3. Set read2 to 94 nt (cell-specific barcodes and UMI)

Include a 6nt read 1 index to ready sublibrary indices

References

1. M. A. Helsby, J. R. Fenn, A. D. Chalmers, Reporting research antibody use: how to increase experimental reproducibility. *F1000Research*. **2**, 153 (2013).
2. C. B. Saper, An open letter to our readers on the use of antibodies. *J. Comp. Neurol.* **493**, 477–478 (2005).
3. S. Picelli *et al.*, Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods*. **10**, 1096–1098 (2013).
4. T. Hashimshony, F. Wagner, N. Sher, I. Yanai, CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell Rep.* **2**, 666–673 (2012).
5. D. A. Jaitin *et al.*, Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science*. **343**, 776–9 (2014).
6. E. Z. Macosko *et al.*, Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*. **161**, 1202–1214 (2015).
7. A. M. Klein *et al.*, Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*. **161**, 1187–201 (2015).
8. T. M. Gierahn *et al.*, Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat. Methods*. **14**, 395–398 (2017).
9. G. X. Y. Zheng *et al.*, Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
10. B. Tasic *et al.*, Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.* **19**, 335–46 (2016).
11. A. Zeisel *et al.*, Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*. **347**, 1138–42 (2015).
12. S. Marques *et al.*, Oligodendrocyte heterogeneity in the mouse juvenile and adult central nervous system. *Science*. **352**, 1326–9 (2016).
13. S. Darmanis *et al.*, A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 7285–90 (2015).
14. B. B. Lake *et al.*, Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science*. **352**, 1586–90 (2016).
15. A. K. Shalek *et al.*, Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*. **498**, 236–40 (2013).
16. D. Grün *et al.*, Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*. **525**, 251–255 (2015).
17. V. Moignard *et al.*, Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat. Biotechnol.* **33**, 269–276 (2015).
18. A. S. Venteicher *et al.*, Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. *Science (80-.).* **355** (2017).
19. I. Tirosh *et al.*, Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*. **352**, 189–96 (2016).
20. L. Sang *et al.*, Control of the reversibility of cellular quiescence by the transcriptional repressor HES1. *Science*. **321**, 1095–100 (2008).
21. C. Zheng *et al.*, Landscape of Infiltrating T Cells in Liver Cancer Revealed by Single-Cell Sequencing. *Cell*. **169**, 1342–1356.e16 (2017).
22. A. B. Rosenberg *et al.*, Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*. **360**, 176–182 (2018).
23. B. Lacar *et al.*, Nuclear RNA-seq of single neurons reveals molecular signatures of activation. *Nat. Commun.* **7**, 11022 (2016).

24. E. R. Thomsen *et al.*, Fixed single-cell transcriptomic characterization of human radial glial diversity. *Nat. Methods*. **13**, 87–93 (2016).
25. Materials and methods are provided as supplementary materials.
26. S. E. Hickman *et al.*, The microglial sensome revealed by direct RNA sequencing. *Nat. Neurosci.* **16**, 1896–905 (2013).
27. O. Matcovitch-Natan *et al.*, Microglia development follows a stepwise program to regulate brain homeostasis. *Science*. **353**, aad8670 (2016).
28. C. Trapnell *et al.*, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–6 (2014).
29. S. W. Levison, J. E. Goldman, Both oligodendrocytes and astrocytes develop from progenitors in the subventricular zone of postnatal rat forebrain. *Neuron*. **10**, 201–12 (1993).
30. E. S. Lein *et al.*, Genome-wide atlas of gene expression in the adult mouse brain. *Nature*. **445**, 168–176 (2007).
31. Allen Institute for Brain Science, Developing Mouse Brain Atlas (2008).
32. T. Iwano, A. Masuda, H. Kiyonari, H. Enomoto, F. Matsuzaki, Prox1 postmitotically defines dentate gyrus cells by specifying granule cell identity over CA3 pyramidal cell fate in the hippocampus. *Development*. **139**, 3051–62 (2012).
33. C. Zhao, W. Deng, F. H. Gage, Mechanisms and Functional Implications of Adult Neurogenesis. *Cell*. **132** (2008), pp. 645–660.
34. A. Lavado, O. V. Lagutin, L. M. L. Chow, S. J. Baker, G. Oliver, Prox1 Is Required for Granule Cell Maturation and Intermediate Progenitor Maintenance During Brain Neurogenesis. *PLoS Biol.* **8**, e1000460 (2010).
35. F. Bonnet *et al.*, Structure and cellular distribution of mouse brain testican. Association with the postsynaptic area of hippocampus pyramidal cells. *J. Biol. Chem.* **271**, 4373–80 (1996).
36. S. Herculano-Houzel, The human brain in numbers: a linearly scaled-up primate brain. *Front. Hum. Neurosci.* **3**, 31 (2009).
37. K. Nakashima, H. Umeshima, M. Kengaku, Cerebellar granule cells are predominantly generated by terminal symmetric divisions of granule cell precursors. *Dev. Dyn.* **244**, 748–758 (2015).
38. A. Sudarov, A. L. Joyner, Cerebellum morphogenesis: the foliation pattern is orchestrated by multi-cellular anchoring centers. *Neural Dev.* **2**, 26 (2007).
39. J. C. Chang *et al.*, Mitotic Events in Cerebellar Granule Progenitor Cells That Expand Cerebellar Surface Area Are Critical for Normal Cerebellar Cortical Lamination in Mice. *J. Neuropathol. Exp. Neurol.* **74**, 261–272 (2015).
40. K. Schilling, J. Oberdick, F. Rossi, S. L. Baader, Besides Purkinje cells and granule neurons: an appraisal of the cell biology of the interneurons of the cerebellar cortex. *Histochem. Cell Biol.* **130**, 601–615 (2008).
41. J. Altman, S. A. Bayer, Development of the precerebellar nuclei in the rat: I. The precerebellar neuroepithelium of the rhombencephalon. *J. Comp. Neurol.* **257**, 477–489 (1987).
42. S. M. Maricich, K. Herrup, Pax-2 expression defines a subset of GABAergic interneurons and their precursors in the developing murine cerebellum. *J. Neurobiol.* **41**, 281–94 (1999).
43. G. Weisheit *et al.*, Postnatal development of the murine cerebellar cortex: formation and early dispersal of basket, stellate and Golgi neurons. *Eur. J. Neurosci.* **24**, 466–478 (2006).
44. Y. L. Petracca *et al.*, The late and dual origin of cerebrospinal fluid-contacting neurons in the mouse spinal cord. *Development*. **143**, 880–891 (2016).
45. A. Enjin *et al.*, Identification of novel spinal cholinergic genetic subtypes disclose Chodl and Pitx2 as markers for fast motor neurons and partition cells. *J. Comp. Neurol.* **518**, 2284–2304 (2010).
46. M. Lalancette-Hebert, A. Sharma, A. K. Lyashchenko, N. A. Shneider, Gamma motor neurons survive and exacerbate alpha motor neuron degeneration in ALS. *Proc. Natl. Acad. Sci. U. S. A.*

113, E8316–E8325 (2016).
47. Allen Institute for Brain Science, Allen Mouse Spinal Cord Atlas (2008).
48. A. Subramanian *et al.*, A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell*. **171**, 1437–1452.e17 (2017).
49. W. Saelens, R. Cannoodt, Y. Saeys, A comprehensive evaluation of module detection methods for gene expression data. *Nat. Commun.* **9**, 1090 (2018).
50. E. Rosati *et al.*, Overview of methodologies for T-cell receptor repertoire analysis. *BMC Biotechnol.* **17**, 61 (2017).
51. G. Yaari, S. H. Kleinstein, Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Med.* **7**, 121 (2015).
52. M. Nagahashi *et al.*, Next generation sequencing-based gene panel tests for the management of solid tumors. *Cancer Sci.* **110**, 6 (2019).
53. L. Tafe *et al.*, P3.09-18 Identification of MET exon 14 Skipping Mutations by FusionPlexTM Solid Tumor Panel. *J. Thorac. Oncol.* **13**, S954–S955 (2018).
54. I. Garcia-Murillas *et al.*, *Sci. Transl. Med.*, in press, doi:10.1126/scitranslmed.aab0021.
55. D. M. Betters, Use of Flow Cytometry in Clinical Practice. *J. Adv. Pract. Oncol.* **6**, 435 (2015).
56. J. Bonnevier, C. Hammerbeck, C. Goetz, (Springer, Cham, 2018), pp. 1–11.
57. C. S. Carlson *et al.*, Using synthetic templates to design an unbiased multiplex PCR assay. *Nat. Commun.* **4**, 2680 (2013).
58. E. Kalle, M. Kubista, C. Rensing, Multi-template polymerase chain reaction. *Biomol. Detect. Quantif.* **2**, 11–29 (2014).
59. A. Rodríguez, M. Rodríguez, J. J. Córdoba, M. J. Andrade, (Humana Press, New York, NY, 2015), pp. 31–56.
60. F. Mertes *et al.*, Targeted enrichment of genomic DNA regions for next-generation sequencing. *Brief. Funct. Genomics*. **10**, 374–86 (2011).
61. J. Dapprich *et al.*, The next generation of target capture technologies - large DNA fragment enrichment and sequencing determines regional genomic variation of high complexity. *BMC Genomics*. **17**, 486 (2016).
62. J. St. John, T. W. Quinn, Rapid capture of DNA targets. *Biotechniques*. **44**, 259–264 (2008).
63. M. Oobatake, S. Takahashi, T. Ooi, Conformational stability of ribonuclease T1. II. Salt-induced renaturation. *J. Biochem.* **86**, 65–70 (1979).
64. C. N. Pace, G. R. Grimsley, J. A. Thomson, B. J. Barnett, Conformational stability and activity of ribonuclease T1 with zero, one, and two intact disulfide bonds. *J. Biol. Chem.* **263**, 11820–5 (1988).
65. A. D. Gelinas, D. R. Davies, N. Janjic, Embracing proteins: structural themes in aptamer–protein complexes. *Curr. Opin. Struct. Biol.* **36**, 122–132 (2016).
66. M. Shugay *et al.*, VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Res.* **46**, D419–D427 (2018).
67. P. Simon *et al.*, Functional TCR Retrieval from Single Antigen-Specific Human T Cells Reveals Multiple Novel Epitopes. *Cancer Immunol Res*. **2** (2014), doi:10.1158/2326-6066.CIR-14-0108.
68. E. T. Wang *et al.*, Alternative isoform regulation in human tissue transcriptomes. *Nature*. **456**, 470–6 (2008).
69. L. Cartegni, M. L. Hastings, J. A. Calarco, E. de Stanchina, A. R. Krainer, Determinants of exon 7 splicing in the spinal muscular atrophy genes, SMN1 and SMN2. *Am. J. Hum. Genet.* **78**, 63–77 (2006).
70. T. Sterne-Weiler, J. R. Sanford, Exon identity crisis: disease-causing mutations that disrupt the splicing code. *Genome Biol.* **15**, 201 (2014).
71. K. Sathasivam *et al.*, Aberrant splicing of HTT generates the pathogenic exon 1 protein in

Huntington disease. *Proc. Natl. Acad. Sci.* **110**, 2366–2370 (2013).
72. I. Vorechovský, Aberrant 3' splice sites in human disease genes: mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization. *Nucleic Acids Res.* **34**, 4630–41 (2006).
73. Y. Lee, D. C. Rio, Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annu. Rev. Biochem.* **84**, 291–323 (2015).
74. A. B. Rosenberg *et al.*, Learning the Sequence Determinants of Alternative Splicing from Millions of Random Sequences Article Learning the Sequence Determinants of Alternative Splicing from Millions of Random Sequences. *Cell*. **163**, 698–711 (2015).
75. G. Singh, T. a. Cooper, Minigene reporter for identification and analysis of cis elements and trans factors affecting pre-mRNA splicing. *Biotechniques*. **41**, 177–181 (2006).
76. D. Di Giacomo *et al.*, Functional analysis of a large set of BRCA2 exon 7 variants highlights the predictive value of hexamer scores in detecting alterations of exonic splicing regulatory elements. *Hum. Mutat.* **34**, 1547–57 (2013).
77. T. A. Cooper, Use of minigene systems to dissect alternative splicing elements. *Methods*. **37**, 331–40 (2005).
78. X. Roca, A. R. Krainer, I. C. Eperon, Pick one, but be quick: 5' splice sites and the problems of too many choices. *Genes Dev.* **27**, 129–44 (2013).
79. M. Ruggiu *et al.*, A role for SMN exon 7 splicing in the selective vulnerability of motor neurons in spinal muscular atrophy. *Mol. Cell. Biol.* **32**, 126–38 (2012).
80. F. Pagani, M. Raponi, F. E. Baralle, Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 6368–72 (2005).
81. F. Pagani, New type of disease causing mutations: the example of the composite exonic regulatory elements of splicing in CFTR exon 12. *Hum. Mol. Genet.* **12**, 1111–1120 (2003).
82. P. Momeni *et al.*, Clinical and pathological features of an Alzheimer's disease patient with the MAPT Delta K280 mutation. *Neurobiol. Aging*. **30**, 388–93 (2009).
83. R. Kole, A. R. Krainer, S. Altman, RNA therapeutics: beyond RNA interference and antisense oligonucleotides. *Nat. Rev. Drug Discov.* **11**, 125–40 (2012).
84. K. A. McQuisten, A. S. Peek, Identification of sequence motifs significantly associated with antisense activity. *BMC Bioinformatics*. **8**, 184 (2007).
85. T. A. Vickers *et al.*, Efficient reduction of target RNAs by small interfering RNA and RNase H-dependent antisense agents. A comparative analysis. *J. Biol. Chem.* **278**, 7108–18 (2003).
86. D. Baralle, M. Baralle, Splicing in action: assessing disease causing sequence changes. *J. Med. Genet.* **42**, 737–48 (2005).
87. Y. Hua, T. a. Vickers, B. F. Baker, C. F. Bennett, A. R. Krainer, Enhancement of SMN2 exon 7 inclusion by antisense oligonucleotides targeting the exon. *PLoS Biol.* **5**, 729–744 (2007).
88. R. Kole, B. J. Leppert, Targeting mRNA Splicing as a Potential Treatment for Duchenne Muscular Dystrophy. *Discov. Med.* **14**, 59–69 (2012).
89. M. J. Yang *et al.*, Mitral and Tufted Cells Are Potential Cellular Targets of Nitration in the Olfactory Bulb of Aged Mice. *PLoS One*. **8**, e59673 (2013).
90. Y. I. Kawasawa *et al.*, RNA-seq analysis of developing olfactory bulb projection neurons. *Mol. Cell. Neurosci.* **74**, 78–86 (2016).
91. J. P. Doyle *et al.*, Application of a translational profiling approach for the comparative analysis of CNS cell types. *Cell*. **135**, 749–62 (2008).
92. E. A. Alcamo *et al.*, Satb2 Regulates Callosal Projection Neuron Identity in the Developing Cerebral Cortex. *Neuron*. **57**, 364–377 (2008).
93. K. Hesse *et al.*, AP-2δ Is a Crucial Transcriptional Regulator of the Posterior Midbrain. *PLoS One*. **6**, e23483 (2011).

94. H. Valcanis, S.-S. Tan, J. G. Parnavelas, Layer specification of transplanted interneurons in developing mouse neocortex. *J. Neurosci.* **23**, 5113–22 (2003).
95. M. Ogawa *et al.*, The reeler gene-associated antigen on cajal-retzius neurons is a crucial molecule for laminar organization of cortical neurons. *Neuron*. **14**, 899–912 (1995).
96. L. He *et al.*, Analysis of the brain mural cell transcriptome. *Sci. Rep.* **6**, 35108 (2016).
97. R. Satija, Cost Per Cell, (available at https://satijalab.org/costpercell).
98. J. Baran-Gale, T. Chandra, K. Kirschner, Experimental design for single-cell RNA sequencing. *Brief. Funct. Genomics*. **17**, 233–239 (2018).
99. D. G. Gibson, Enzymatic Assembly of Overlapping DNA Fragments. *Methods Enzymol.* **498**, 349–361 (2011).
100. A. Dobin *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. **29**, 15–21 (2013).
101. E. Zorita, P. Cuscó, G. J. Filion, Starcode: sequence clustering based on all-pairs search. *Bioinformatics*. **31**, 1913–9 (2015).
102. Z. Yao *et al.*, A Single-Cell Roadmap of Lineage Bifurcation in Human ESC Models of Embryonic Brain Development. *Cell Stem Cell*. **20**, 120–134 (2017).
103. K. Shekhar *et al.*, Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. *Cell*. **166**, 1308–1323.e30 (2016).
104. N. Habib *et al.*, Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. *Science*. **353**, 925–8 (2016).
105. F. Pedregosa *et al.*, Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
106. L. van der Maaten, Accelerating t-SNE using Tree-Based Algorithms. *J. Mach. Learn. Res.* **15**, 3221–3245 (2014).
107. M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. *Proc. Second Int. Conf. Knowl. Discov. Data Min.* (1996), pp. 226–231.
108. X. Qiu *et al.*, Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* (2017), doi:10.1038/nmeth.4402.
109. D. A. Bolotin *et al.*, MiXCR: software for comprehensive adaptive immunity profiling. *Nat. Methods*. **12**, 380–381 (2015).
110. Y.-C. Chiu *et al.*, Foxp2 regulates neuronal differentiation and neuronal subtype specification. *Dev. Neurobiol.* **74**, 723–738 (2014).
111. T. Miyashita *et al.*, Neurotrophin-3 Is Involved in the Formation of Apical Dendritic Bundles in Cortical Layer 2 of the Rat. *Cereb. Cortex*. **20**, 229–240 (2010).
112. T. G. Belgard *et al.*, A Transcriptomic Atlas of Mouse Neocortical Layers. *Neuron*. **71**, 605–616 (2011).
113. A. Nagalski *et al.*, Molecular anatomy of the thalamic complex and the underlying transcription factors. *Brain Struct. Funct.* **221**, 2493–2510 (2016).
114. M. B. Wisniewska *et al.*, LEF1/beta-catenin complex regulates transcription of the Cav3.1 calcium channel gene (Cacna1g) in thalamic neurons of the adult brain. *J. Neurosci.* **30**, 4957–69 (2010).
115. H. Song *et al.*, Ascl1 and Helt act combinatorially to specify thalamic neuronal identity by repressing Dlxs activation. *Dev. Biol.* **398**, 280–291 (2015).
116. J. D. Corrales, S. Blaess, E. M. Mahoney, A. L. Joyner, The level of sonic hedgehog signaling regulates the complexity of cerebellar foliation. *Development*. **133**, 1811–21 (2006).
117. E. Salero, M. E. Hatten, Differentiation of ES cells into cerebellar neurons. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 2997–3002 (2007).
118. C. A. D'Souza *et al.*, Identification of a set of genes showing regionally enriched expression in the mouse brain. *BMC Neurosci.* **9**, 66 (2008).

119. N. X. Tritsch, J. B. Ding, B. L. Sabatini, Dopaminergic neurons inhibit striatal output through non-canonical release of GABA. *Nature*. **490**, 262–6 (2012).
120. M. Moriyama *et al.*, Complement receptor 2 is expressed in neural progenitor cells and regulates adult hippocampal neurogenesis. *J. Neurosci.* **31**, 3981–9 (2011).
121. J. V. Nielsen, J. B. Blom, J. Noraberg, N. A. Jensen, Zbtb20-Induced CA1 Pyramidal Neuron Development and Area Enlargement in the Cerebral Midline Cortex of Mice. *Cereb. Cortex*. **20**, 1904–1914 (2010).
122. M. S. Cembrowski, L. Wang, K. Sugino, B. C. Shields, N. Spruston, Hipposeq: a comprehensive RNA-seq database of gene expression in hippocampal principal neurons. *Elife*. **5**, e14997 (2016).
123. C. L. Thompson *et al.*, A High-Resolution Spatiotemporal Atlas of Gene Expression of the Developing Mouse Brain. *Neuron*. **83**, 309–323 (2014).
124. Z. Xie *et al.*, Zbtb20 is essential for the specification of CA1 field identity in the developing hippocampus. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 6510–5 (2010).
125. H. M. Knight *et al.*, GRIK4/KA1 protein expression in human brain and correlation with bipolar disorder risk variant status. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* **159B**, 21–29 (2012).
126. S. Shah, E. Lubeck, W. Zhou, L. Cai, In Situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus. *Neuron*. **92**, 342–357 (2016).
127. R. P. Seal *et al.*, Injury-induced mechanical hypersensitivity requires C-low threshold mechanoreceptors. *Nature*. **462**, 651–655 (2009).
128. P. Alifragis, A. Liapi, J. G. Parnavelas, Lhx6 Regulates the Migration of Cortical Interneurons from the Ventral Telencephalon But Does Not Specify their GABA Phenotype. *J. Neurosci.* **24**, 5643–5648 (2004).
129. E. Sánchez-Mendoza *et al.*, Review: Could neurotransmitters influence neurogenesis and neurorepair after stroke? *Neuropathol. Appl. Neurobiol.* **39**, 722–735 (2013).
130. H. Abraham, C. G. Pérez-García, G. Meyer, p73 and Reelin in Cajal-Retzius Cells of the Developing Human Hippocampal Formation. *Cereb. Cortex*. **14**, 484–495 (2004).
131. K.-K. Cheung, S. C. Mok, P. Rezaie, W. Chan, Dynamic expression of Dab2 in the mouse embryonic central nervous system. *BMC Dev. Biol.* **8**, 76 (2008).
132. O. Butovsky *et al.*, Identification of a unique TGF-β-dependent molecular and functional signature in microglia. *Nat. Neurosci.* **17**, 131–43 (2014).
133. J. T. Ehmsen *et al.*, The astrocytic transporter SLC7A10 (Asc-1) mediates glycinergic inhibition of spinal cord motor neurons. *Sci. Rep.* **6**, 35592 (2016).
134. S. Chuikov, B. P. Levi, M. L. Smith, S. J. Morrison, Prdm16 promotes stem cell maintenance in multiple tissues, partly by regulating oxidative stress. *Nat. Cell Biol.* **12**, 999–1006 (2010).
135. Z. H. Afsari, W. M. Renno, E. Abd-el-basset, Alteration of Glial Fibrillary Acidic Proteins Immunoreactivity in Astrocytes of the Spinal Cord Diabetic Rats. *Anat. Rec. Adv. Integr. Anat. Evol. Biol.* **291**, 390–399 (2008).
136. A. S. Saab *et al.*, Bergmann glial AMPA receptors are required for fine motor coordination. *Science*. **337**, 749–53 (2012).
137. A. Honoré *et al.*, Isolation, characterization, and genetic profiling of subpopulations of olfactory ensheathing cells from the olfactory bulb. *Glia*. **60**, 404–413 (2012).
138. M. Murthy, S. Bocking, F. Verginelli, S. Stifani, Transcription factor Runx1 inhibits proliferation and promotes developmental maturation in a selected population of inner olfactory nerve layer olfactory ensheathing cells. *Gene*. **540**, 191–200 (2014).
139. J. D. Cahoy *et al.*, A Transcriptome Database for Astrocytes, Neurons, and Oligodendrocytes: A New Resource for Understanding Brain Development and Function. *J. Neurosci.* **28**, 264–278 (2008).

140. M. G. Erlander, A. J. Tobin, The structural and functional heterogeneity of glutamic acid decarboxylase: A review. *Neurochem. Res.* **16**, 215–226.
141. P. R. Brumovsky, P. R., VGLUTs in Peripheral Neurons and the Spinal Cord: Time for a Review. *ISRN Neurol.* **2013**, 829753 (2013).