Modeling the Transcriptional Landscape of *in vitro* Neuronal Differentiation and ALS Disease

Elena Konstantinovna Kandror

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2019

ABSTRACT

Modeling the Transcriptional Landscape of *in vitro* Neuronal Differentiation and ALS Disease

Elena Konstantinovna Kandror

The spinal cord is a complex structure responsible for processing sensory inputs and motor outputs. As such, the developmental and spatial organization of cells is highly organized. Diseases affecting the spinal cord, such as Amyotrophic Lateral Sclerosis (ALS), result in the disruption of normal cellular function and intercellular interactions, culminating in neurodegeneration. Deciphering disease mechanisms requires a fundamental understanding of both the normal development of cells within the spinal cord as well as the homeostatic environment that allows for proper function. Biological processes such as cellular differentiation, maturation, and disease progression proceed in an asynchronous and cell type-specific manner. Until recently, bulk measurements of a mixed population of cells have been key in understanding these events. However, bulk measurements can obscure the molecular mechanisms involved in branched or coinciding processes, such as differential transcriptional responses occurring between subpopulations of cells. Measurements in individual cells have largely been restricted to 4 color immunofluorescence assays, which provide a solid but limited view of molecular-level changes. Recently, developments in single cell RNA-sequencing (scRNA-Seq) have provided an avenue of accurately profiling the RNA expression levels of thousands of genes concomitantly in an individual cell. With this increased experimental precision comes the ability to explore pathways that are differentially activated in subpopulations of cells, and to determine the transcriptional programs that underlie complex biological processes. In this dissertation, I will first review the key features of scRNA-Seq and downstream analysis. I will then discuss two applications of scRNA-seq: 1) the *in vitro* differentiation of mouse embryonic stem cells into motor neurons, and 2) the effect of the ALS-associated gene $SOD1^{G93A}$ expression on cultured motor neurons in a cellular model of ALS.

Contents

List of Figures and Tables

Figures

. . . . . .

Tables

Acknowledgements

Tom Maniatis, for letting me be a part of his lab. Thank you for your guidance, support, and mentorship as I learned how to work as a scientist.

Hynek Wichterle, for serving as the chair of my thesis committee; George Mentis, Neil Shneider, Stavros Lomvardas, and Ai Yamamoto for serving on my thesis committee and thesis defense committee. Thank you for your support and helpful discussions, your guidance, your suggestions, and for sharing your knowledge with me.

Raul Rabadan and Pablo Camara, for collaboration on the scTDA project. Thank you for your insights and patience as I learned about theory, analysis, and computation.

Rina Leah Davidson, Lexi Peterson, and Tom Roberts for experimental support. Thank you for your hard work, and for always being A team players.

Ira Schieren and Michael Kissner, for help with flow cytometry.

Members of the Maniatis lab and others, for interesting discussions.

Zaia Sivo, for keeping me on track.

Abbas Rizvi, for being the strongest possible mentor, guide, advocate, and role model. None of this work could have been possible without your thought, foresight, and effort. Thank you for teaching me how to think critically, to do experiments with awareness, and to work through the hard times. I never knew that “out of compassion for the science” was an important reason to work diligently, but now I do.

I have the blessing of a caring and mentoring family. With the understanding and support that comes from my parents and sister, all of who know the landscape of academic science, I have been able to go through this process without ever feeling like I was alone. No matter what I struggled through, they never wavered in their belief in me. I would like to thank Mama, for the phone calls and care packages that kept me going through the most difficult times and made me feel human, comforted, and safe; Papa, for his guidance and firm confidence that I was capable and doing the right thing; Vera, for being the absolute best sister, advocate, and role model, for always keeping her heart open to me, and consistently helping me feel loved and sane. I would also like to thank Sasha and Sofia, my nieces, for being such funny and charming girls; being with them is delightful, and I look forward to their company even when I feel like I am drowning with work. Finally, I would like to thank my grandparents, who helped shape me; they taught me the multiplication table, told me stories, and gave me so much love. To my family: I love you. Thank you.

Dedication

I am dedicating this to Abbas Rizvi, who saw me struggling as a graduate student and made it his mission to make sure I could succeed. It is not just that none of the work in this thesis would have been possible without his herculean efforts. If not for his intervention, I would have gone through science hopelessly demoralized, without confidence, gaining little and giving less back. He has given me more than I can put into words, and I will forever be grateful that he had the compassion and the patience to be my teacher.

Introduction

The spinal cord is responsible for processing somatosensory inputs and motor output. It is a symmetric structure; the organization of the left and right halves of the spinal cord are largely equivalent, and respectively innervate the left and right half of the body. During development, highly organized circuits compromised of both neuronal and glial cells are generated and compartmentalized into discrete regions of the spinal cord (Figure 1). The positions of these regions, defined rostro-caudally by spinal vertebrae and dorso-ventrally by Rexed laminae, contain the cell bodies of neurons that are responsible for information processing to and from distinct parts of the body. Lower motor neurons, located in the anterior grey column (lamine VIII and IX), extend axons out of the spinal cord and directly innervate muscles. Forelimbs are innervated by motor neurons located in the cervical segments of the spinal cord, axial muscles by the thoracic segments, and hindlimbs by the lumbar segments. Within these segments, motor neurons are further organized medio-laterally into motor columns, which are responsible for innervating morphogenically similar muscle groups. During spinal cord development, the specification of cell state and spatial location is dictated by the interplay of tightly controlled molecular cues and the transcriptional programs they regulate[1]. Within the notochord, the floorplate and the roofplate establish a dorsal-ventral gradient of Sonic Hedgehog (Shh) and an opposing gradient of Bone Morphogenic Protein (BMP). Cells experiencing dorsal-ventral levels of Shh above a particular threshold activate type I homeodomain genes and repress the transcription of type II homeodomain genes[2]. As a result, ventral boundaries are established, with some cells expressing type I or type II homeodomain transcription factors. In addition, boundaries between these cells are enhanced by the cross repressive actions of these two classes of homeodomain transcription factors. In a similar fashion, the opposing BMP gradient acts in a dorsal to ventral fashion, ensuring sharp transcriptional domains within the developing spinal cord. MNs arise from a region expressing Class I Nkx6.1 in the absence of Irx3 or Dbx2. In addition, within this domain, class II Pax6 is expressed in the absence of Nkx2.2 and Nkx2.9. This combination of transcription factors creates a system of de-repression, where pan-neuronal

genes are activated. However, those genes that specify neuronal identity other than MNs are repressed by the presence of Nkx6.1 and Pax6[3]. Progenitors within this region then express Olig2, a transcription factor responsible for the transcriptional activation of MN specific genes, such as the transcription factors Hb9, Mnr2, Lim3 and Isl1/2[4]. Markers used to monitor the maturity of MNs are limited, but include the aforementioned transcription factor, Hb9, and choline acetyltransferase (ChAT), with Hb9 expressed early and ChAT expressed late[5]. Motor columns can be identified by the expression of unique combinations of transcription factors and cell surface markers[6]. Each motor column contains motor pools, which are a collection of motor neurons that innervate a single muscle. Each motor pool consists of a combination of motor neuron subtypes: Alpha (α) motor neurons that innervate the extrafusal fibers (standard skeletal muscle), and beta (β) and gamma (γ) motor neurons that innervate the intrafusal fibers (muscle spindle). α motor neurons can further be classified based on the type of muscle they innervate: fast-twitch fatigable (FF) motor neurons have the largest diameter and innervate large muscles, fast-twitch fatigue-resistant (FR) are intermediate size, and slow-twitch fatigue-resistant (S) are the smallest[6].

Given the precise organization of motor neurons within the spinal cord, disruptions of any part of the motor circuit result in impaired motor function. Partial or complete paralysis can result from a traumatic spinal cord injury, exposure to venom, bacterial and viral infections, autoimmune and developmental disorders, and neurodegenerative diseases. The manner in which these occurrences impair motor neuron function are highly variable, as is their prognosis. In cases of temporary paralysis, no irreversible damage to the motor neuron occurs. For example, in incomplete spinal cord injuries, locomotion may be regained after pressure and inflammation are alleviated from the point of trauma. As delineated above, the numbers, positions, and identities of motor neurons in the spinal cord are established during early development and maintained throughout adulthood. Consequently, clinical cases involving degeneration of motor neurons such as Amyotrophic Lateral Sclerosis (ALS) are linked with a much worse prognosis.

ALS is the most common form of adult-onset motor neuron degenerative disease in the United States. ALS is uniformly fatal, with a life expectancy of 3 – 5 years post diagnosis. 10% of cases are familial, while 90% are sporadic[7]. Mutations in over 20 genes have been associated with both forms of ALS[8]. Symptoms start with muscle weakness in a focal point, often in the arms or legs (called limb-onset ALS), which then progresses to paralysis in all voluntary muscles in the body. Patient death is due to respiratory failure, which occurs when muscles controlling breathing become affected. While the symptoms of ALS result from loss of muscle function, the onset and spread of the disease are due to motor neuron degeneration in the central nervous system. In the spinal cord, motor neuron death proceeds in a stereotyped manner, with motor neuron loss being most severe in the motor pools innervating muscles at the site of ALS onset. FF α motor neurons within motor pools are selectively vulnerable to degeneration, while β and γ motor neurons are resistant[9]. ALS spreads first to neighboring motor pools in a motor column, then contiguously propagates both ipsilaterally along the rostrocaudal axis and contralaterally to the complementary motor column on the opposite side of the spinal cord. In the cortex, upper motor neurons also contribute to disease progression. Upper motor neuron loss is again most severe in the region responsible for innervating the site of ALS onset, spreading radially out towards adjacent regions in the motor cortex. The degree of upper versus lower motor neuron involvement is variable between patients, as is the clinical manifestation of paralysis progression through different muscles groups.

While motor neurons are selectively vulnerable to degeneration in ALS, pathology also occurs in neighboring glial cells[10]. This is consistent with the observation that genes associated with familial and sporadic ALS are expressed in both neuronal and glial populations. Much of what is known about the cell type-specific contributions to, and underlying molecular mechanisms of, disease progression come from *in vivo* and *in vitro* models based on mutations in these genes. The most studied of these is superoxide dismutase 1 (SOD1). Introducing an ALS-associated variant of SOD1, such as $SOD1^{G93A}$, into model organisms results in a phenocopy of ALS disease[11]. *In vivo* mouse models have revealed the relative contribution of glia and neurons to ALS onset, rate of

progression, and lifespan through restricted expression of SOD1^{G93A} to particular cell types[12-16]. Astrocytes, which normally function as support cells to neurons, become reactive: they lose their beneficial functions and begin to secrete toxic factors into the microenvironment, inducing and accelerating motor neuron degeneration[17-20]. Microglia, the immune cells of the central nervous system, become activated and release pro-inflammatory signaling molecules[21-23]. Mature oligodendrocytes are replaced by an immature population lacking the capacity to myelinate axons[24, 25]. This complicated milieu of pathological events can be modeled more simply with *in vitro* cultures of defined cells. *In vitro* models can be generated from primary cells isolated from animal models or from human patients. Primary astrocytes, microglia, and cortical neurons can be expanded in culture and studied directly, while motor neurons are commonly obtained through reprogramming stem cells. Despite introducing artificiality into the system, *in vitro* models have the benefit of being scalable, reproducible, and easily testable.

in vitro models use glia-motor neuron cocultures to better replicate the endogenous spinal cord environment. Methods have been established for isolating viable astrocytes from the human[26] and murine[27] cortex, microglia from the murine cortex and spinal cord[28] and postmortem human brain[29], and oligodendrocytes from the rat optic nerve[30]. Spinal motor neuron purification and culture protocols have also been established for cells from embryonic[31, 32] and adult mice[33], however the yield of motor neurons is limiting. Successful differentiation of motor neurons (ESMNs) from murine embryonic stem cells (mESCs) has allowed for large-scale studies of ALS[34]. mESC lines are generated from early stage blastocysts that are cultured, expanded, and propagated[35]. mESCs retain the properties of the blastocyst, so genetic modifications in the blastocyst such as inclusion of a fluorescent reporter or insertion of a mutated copy of a gene will be retained in the mESC. In context of ALS, mESC lines have been generated from blastocysts that have a motor neuron specific promoter Hb9-driven eGFP cassette[34], which allows for FACS purification of differentiated motor neurons, and expression of a human ALS-associated gene such as SOD1^{G93A},[36] TDP43^{A315T}, FUS^{R514S}, FUS^{R521C,} FUS^{P525L},[37] or FUS^{P517L} [38]. These studies have revealed the non-cell autonomous effects of astrocytes on motor neuron degeneration,

transcriptional pathways that are dysregulated in a cell type specific way in ALS, protein aggregation, mislocalization, and phenotypic changes such as axonal breakage and excitotoxicity.

Differentiation protocols for human embryonic stem cells (hESCs), and human induced pluripotent stem cells (hiPSCs) to motor neurons have also been established[39, 40]. These protocols rely on activating the endogenous signaling pathways employed during morphogenesis to drive stem cells to differentiate towards a particular lineage[41, 42]. Alternatively, patient fibroblasts can be directly reprogrammed into motor neurons without going through a stem cell-like state[43]. A key benefit of these models is that they are patient-specific, and can therefore be generated from sporadic patients and carry the genetic background that could predispose an individual to develop ALS. hiMNs and hESMNs show may common pathological features of ALS, however often (but not always[44]) remain equally viable to control iMNs under non-cell autonomous stress from ALS astrocytes[45-48]. Stem cell-derived motor neurons not harboring an ALS mutation have also been used to study the impact of ALS patient-derived astrocytes and primary murine astrocytes on motor neuron degeneration[49-52]. Fully humanized *in vitro* models have also provided huge advances in understanding ALS. For example, necroptosis was implicated as the driving factor behind motor neuron degeneration[53]. While these studies have led to a number of important discoveries in ALS, there are several biological complications associated with them that can convolute results and obscure relevant information.

First, iPSCs and directly reprogrammed iMNs retain epigenetic signatures of the parent cell they were taken from[54, 55]. The extent to which these signatures are retained between individual cells of a reprogramming protocol has not been fully explored, and can account for variability in disease models. Furthermore, the efficiency of motor neuron differentiation varies even within replicates of the same protocol, and the heterogeneity of resulting post-mitotic populations has not been well characterized. The cells that result from motor neuron differentiations have characteristic phenotypic properties including stereotyped electrophysiological responses, ability

to form neuromuscular junctions with muscles, and expression of marker genes including Hb9, ChAT, and Isl1[56]. Transplantation experiments have shown that these cells can be grafted into chick spinal cords where they assemble into appropriate motor columns[57]. The rostro-caudal patterning of cervical, thoracic and lumbar motor neurons can be observed *in vitro* through expression of Hox genes. A common reagent in neuronal differentiations, retinoic acid[58], has been shown to result in motor neurons with a cervical identity[34, 59]. Protocols abstaining from retinoic acid induction result in a lower efficiency differentiation of motor neurons with more rostral identity[60]. *In vivo* studies have shown that limb-innervating motor neurons in the lateral motor column are $Foxp1^{+}/Lhx3^{-}$, and iMNs and ESMNs can be generated to reflect this transcriptional program[41]. However, parallel transcriptional studies have shown that human iPSC-derived neurons have a transcriptional program that overall most closely resembles immature embryonic stages of development[61]. A comparative study of 7 week old hiMNs and hESMNs with fetal spinal cord, adult spinal cord, and laser capture microdissected adult spinal motor neurons also found the strongest transcriptional correlation between fetal tissue and *in vitro* cells[62]. Positive selection for Hb9::GFP positive hESMNs[63] did not improve the age-related differences. Neuronal maturation can be measured via several methods, including cell morphology and electrophysiology. In a timecourse study on Hb9::GFP positive hESMNs, cultured cells displayed more mature electrophysiological properties at later timepoints, however the variability between cells also increased over time[64].

Commonalities between cells arising from motor neuron differentiations have been established and standards for defining a successful motor neuron differentiation have been implemented. These characteristic transcriptional, electrophysiological, and morphological properties are important for controlled experiments, but can belie the heterogeneity in the post-mitotic population. As a measurement, the bulk transcriptional profile of ESMNs and iMNs encompasses a broad definition of motor neurons, but until high resolution profiling of single cells is achieved, these bulk measurements conceal individuality and may obscure ALS-associated changes that occur in subsets of cells. As discussed above, spinal motor neurons are a diverse population that

are differentially susceptible to degeneration in ALS. Some of this heterogeneity is propagated in *in vitro* models.

Motor neuron differentiations are variable in their efficiency, and produce a heterogeneous population of post-mitotic cells. Understanding the dynamics of these differentiations can both shed light on molecular pathways that contribute to neuronal commitment, and help improve modeling of diseases such as ALS. In mouse models, the bulk transcriptional profiles of motor neurons differentiated from mESCs (ESMNs) largely recapitulate the transcriptional profiles of laser-capture microdissected spinal motor neurons[65]. Therefore, the murine model was the starting point for single cell experiments. This, however, belies the heterogeneity of motor neuron populations that make subsets of cells vulnerable or resistant to degeneration during disease. In fact, similar to the progression of ALS in the spinal cord, cultured ESMNs also display asynchronous degeneration that is accelerated by exposure to glia harboring an ALS-associated mutation.

The variability in ALS disease progression both *in vivo* and *in vitro* implies 1) the existence of a stochastic trigger for disease onset, and 2) heterogeneous cellular responses to a common stimulus. Due to the asynchronous progression of ALS and the unique transcriptional responses different cellular subpopulations have to ALS-associated toxicity, bulk measurements taken from a mixed cellular population obscure the signaling pathways activated in individual neurons and glia. Single cell RNA sequencing (scRNA-seq) overcomes this obstacle. Using scRNA-seq, it is possible to measure the transcriptional readouts of individual cells at any timepoint of the disease. scRNA-seq applied over a timecourse of disease development allows the ability to track the progression of transcriptional changes occurring in the system over time.

In bulk sequencing, the transcriptional readout of a homogeneous cellular population can be taken as the equivalent to its identity. Correlation metrics for defining cell types, differential gene expression, and population similarities are based on this foundation[66, 67]. However, it has long

been understood that even a clonal population of cells is not homogeneous, and individuals residing in a population of genetically identical cells have different responses to inductive cues[68]. These differences were shown to be a result of the stochastic, not predetermined, activity of transcriptional templates[69], a process known as translational noise and transcriptional bursting[70, 71]. Random noise in single cell expression has been found to be critically important in development and cellular diversification[72]. Noise can be measured at a time scale of seconds to hours, and is thought to be of both intrinsic (possibly through chromatin remodeling complexes) and extrinsic (cell size and shape) origin[73]. Stochastic gene expression makes analysis of scRNA-seq datasets challenging because the level of expression of a gene may not relate to its translational readout. Furthermore, in order to calculate accurate similarity between cells, effective differential expression analysis needs to be calculated across multiple connected transcripts. A second complicating factor is transcript dropout. Due to technical limitations in generating single cell libraries, many transcripts that are present in the cell are not captured in the library, leading to artificially inflated levels of zero expression. This is especially true for small RNAs such as miRNAs, long RNAs that are over 2kb in length, and mRNAs that are expressed at low levels. Many technical approaches to single cell sequencing have been established, all of which have some benefits and drawbacks associated with them. Protocols can be categorized by: linear amplification vs nonlinear amplification, plate-based vs droplet-based, and high coverage vs low coverage approaches. As a result of the many variables associated with scRNA-seq, single cell data analysis faces challenges that algorithms developed for bulk RNA sequencing do not need to contend with.

Traditional computational approaches to analyzing scRNA-seq data depend on clustering cells into groups based on the similarity of their transcriptional profiles. These methods fail to account for transitioning cells or contiguous processes. Novel algorithms, such as single cell topological data analysis (scTDA), have been established to delineate transcriptional trajectories that cells go through during a continuous process[74]. scTDA clusters cells in a high dimensional space, allowing for cells to overlap in multiple clusters, before dimensionally reducing the output into a

two dimensional representation. In the representation, each cluster in the high dimension is represented by a single node, and clusters that shared at least one cell in the high dimensional space are connected by a line (Figure 1b). This approach enables statistical analysis of transcriptional progression in the representation, which in turn allows for analysis of pathway activation in subpopulations of cells during the timecourse.

We have applied two different scRNA-seq approaches to two biological questions. First, we aimed to characterize the transcriptional pathways responsible for neurogenic commitment of mESCs during motor neuron differentiation. Second, we examined at the effects the $SOD1^{G93A}$ mutation had on cell autonomous and non-cell autonomous ESMN degeneration *in vitro*. Combining scRNA-seq of ESMN differentiations and *in vitro* models of ALS with scTDA analysis allows for an in depth exploration of the transcriptional landscape surrounding motor neuron identity, heterogeneity, and pathology. It provides a platform for understanding variability in model systems and asynchrony in disease onset and progression. This approach is also a powerful tool that can be applied to other biological systems that exhibit similar convolutions in data, both technical and analytical.

Chapter 1: Single Cell RNA-Sequencing

Introduction: Single Cell RNA Sequencing and Analysis

In 2013, Nature Publishing Group selected single cell sequencing as its Method of the Year, stating that "methods to sequence the DNA and RNA of single cells are poised to transform many areas of biology and medicine"[75]. In 2018, Science Magazine named the application of single cell sequencing to organism development the Breakthrough of the Year[76]. Single cell RNA sequencing (scRNA-seq) allows transcriptome profiling at a fundamental level of resolution, thereby resolving complex system heterogeneity and defining cellular individuality. scRNA-seq has enabled revolutionary discoveries, including the identification of rare cell types[77], molecular rationales for drug resistance[78], and pathways involved in lineage commitment[79].

Broadly, the layout for scRNA-seq experiments is similar to that for bulk RNA-seq, consisting of three steps: 1) reverse transcription of mRNA into complementary DNA (cDNA), 2) amplification of cDNA into a sequencing library, and 3) physical sequencing and analysis. However, technical hurdles – such as the limiting amount of starting material (10pg of RNA in a cell, with only 1-10% of that being mRNA), high variability in RNA content, and large sample sizes necessary for statistical accuracy – a variety of methods exist to perform these steps, all of which have some benefits and disadvantages associated with them. Constant progress is being made on both the experimental and computational fronts of scRNA-seq, allowing for improved gathering and interpretation of data.

scRNA-seq experiments coupled with appropriate computational approaches can yield deep insights into biological processes. A cell's transcriptional and chromatin state reflects both its identity and its dynamic response to perturbations in the environment. As such, scRNA-seq and ATAC-seq is a useful tool for studying cellular identity and development, disease progression and resistance, and other biological events that result in an altered transcriptional state.

Origins of Single Cell RNA Sequencing

The earliest applications of single cell nucleic acid profiling were accomplished not by sequencing, but by *in situ* hybridization in 1968[80]. The molecular foundation of these imaging experiments – that a melted DNA molecule can reassociate with a complementary DNA or RNA sequence – laid the groundwork for all future molecular biology research. While *in situ* hybridizations were limited to probing for a single gene or transcript of interest, they clearly demonstrated that even within a highly similar population of cells there existed transcriptional individuality, and that discrepancies in morphologically analogous cells were vital for function[81-83]. In 1976, the first full length *in vitro* synthesis of cDNA from mRNA transcripts was achieved, opening the doors to study specific gene expression in cells[84, 85]. Almost 20 years later, in 1992, the first quantitative single cell RNA profiling studies were published, looking at 8 genes across less than 20 cells[86-88]. Already, two experimental approaches tailored towards two biological applications were enacted, the basis of which persists in current scRNA-seq protocols: linear and non-linear amplification schemes. After cDNA synthesis, linear amplification uses T7 *in vitro* transcription to generate amplified RNA (aRNA), which then is converted back to cDNA for data generation[89, 90]. This process is more time consuming but introduces less bias into the final data, and was purposed towards quantifying the relative abundance of mRNAs in cells. Non-linear amplification uses a PCR primer to directly amplify cDNA[91] which, while faster, can introduce artifacts into the data, absent the use of unique molecular identifiers, that account for PCR based artifacts. This approach was purposed towards determining which subunits of the AMPA receptor were expressed in a single cell, and for discovering a family of pheromone receptor genes[92]. The analysis of both these experiments was simple: the abundance of a transcript was directly proportional to the intensity of its corresponding band on an agarose gel or Southern Blot.

The field then moved towards higher resolution experiments with the development of pipelines that allowed quantification of many transcripts within a single cell[93]. With the invention of laser capture microdissection (LCMD)[94, 95], innovation of cell sorting instruments[96, 97], development of microarrays[98], improvement in the sensitivity of sequencers[99, 100], and generation of reference

genomes and annotations for multiple species and splicing events[101-103], more advanced strategies could be implemented and more information could be generated from single cell experiments. In 2003, the first single cell microarray was published[104], followed in 2009 by the first full-transcriptome scRNA-seq paper[105]. From then on, single cell sequencing technology and methodology have been improved and expanded on[106]. Sequencing costs, the bottleneck for generating large amounts of data, have reduced dramatically. Currently, it is possible to sequence thousands of individual cells at once, detecting thousands of transcripts in each cell. scRNA-seq could not have been accomplished without foundations laid down by the brilliant experiments of early molecular biologists, the productive intersection of industry with academia, and impressive advances in the fields of molecular biology, biochemistry, surface chemistry, microfluidics, optics, and bioinformatics.

Illumina Sequencing Platforms

Due to their accuracy and speed, the most widely used instruments for sequencing libraries are produced by Illumina[106]. These include the benchtop MiSeq and NextSeq 500/550, which provide rapid turnaround of data (run times range from 11 hours for 75bp sequences to 29 hours for 300bp sequences), and the larger production-scale sequencers including the HiSeq2500, HiSeq4000/X, and NovaSeq 6000 (which can take from 13 to 84 hours per run). Apart from accessibility of the instruments – production scale sequencers are often housed in cores, genome centers, or companies, while benchtop sequencers can also be available in research labs – two key user differences are the amount of input material required, and the amount of data generated. Smaller runs generate up to 400 million reads per run, while larger runs can generate up to 20 billion. The number of reads sufficient for an experiment is highly dependent on the number of cells being analyzed, the quality of the library, and the desired number of genes detected. Users furthermore have the ability to specify if the run is single-end or paired-end – whether the library is sequenced from just the 3' or both the 3' and 5' ends – and the length of the output sequences (75bp – 300bp). Furthermore, the sequencing chemistry utilized on HiSeq4000/X and NovaSeq

6000 instruments utilize exAMP approaches, discussed below, which can result in index swapping which can obfuscate results[107].

Illumina incorporates flow cells, a solid-phase sequencing by synthesis platform, in all of its instruments. Adapter sequences flanking the cDNA library anneal to a dense lawn of complementary primers that are grafted onto the surface of the flow cell. The immobilized cDNA undergoes several rounds of PCR on the flow cell, a process called bridge amplification, until microscopic clusters of clonal cDNA are generated. These clusters are the substrates for a modified form of Sanger sequencing know as sequencing by synthesis, in which fluorescently labeled nucleotides are sequentially incorporated into the nascent strand of cDNA. The flow cell is imaged after each incorporation step. The coordinates of consistent and localized fluorescence intensity are registered by the software as clusters, and the fluorescent readout from clusters is converted nucleotide sequences (a process known as “base calling”). The output fastq file contains the sequence of all cDNA clusters registered on the flow cell along with the phred quality score (Q-score) for each base, which reflects the probability that a base call is accurate.

Illumina uses two fabrication designs for flow cells and two approaches to color chemistry for sequencing, so technical considerations should be made in choosing the appropriate instrument for a sequencing run. The HiSeq 4000/X and NovaSeq 6000 use patterned flow cells, in which the P5 and P7 primers are grafted into nanowells in the flow cell instead of being uniformly distributed as a lawn. This, combined with the innovative exclusion amplification (ExAmp) bridge amplification used by these instruments, improves the workable surface area of a flow cell by increasing the density of useable clusters while preventing overclustering or mixed cluster generation from closely annealed cDNAs. Patterned flow cells and ExAmp significantly cut the cost for sequencing, with several drawbacks. First, sequencing runs on these instruments suffer from index swapping: in multiplexed samples, up to 6% of reads from one library can be assigned an incorrect index that came from a different library, effectively contaminating the data[107]. Specialized dual-index primers can be incorporated into the library to compensate for this[107, 108].

Second, the recommended library size distribution for patterned flow cells is much tighter than for traditional flow cells (300bp-500bp, as opposed to up to 1500bp[109]), because ExAmp bridge amplification is restricted to nanowells. This limits the information accessible through protocols that rely on 3' mRNA priming and amplification (discussed later). While all instruments preferentially cluster short cDNA fragments over long ones, the difference is especially pronounced in patterned flow cells. Primer dimers are a common by-product of library generation, and even a 5% carryover of these short fragments into the final library can result in 60% of the reads being overtaken by dimers sequences (unpublished data). Finally, duplication rates in patterned flow cells are higher than traditional ones, due to cross-seeding of cDNA into adjacent nanowells during ExAmp. While duplicated reads are not unique to patterned flow cells, and indeed are universally present in PCR-amplified libraries, they reduce the information content gathered from a sequencing run and can impact downstream differential expression analysis[110].

Low-diversity libraries, in particular scRNA-seq libraries that contain an oligo-dT stretch, form clusters that Illumina software has difficulty distinguishing. This largely stems from confusion in the built-in imaging and analysis platform, which attempts to identify clusters based on localized and synchronized fluorescence on the flow cell. When adjacent clusters continuously incorporate the same fluorescent base, and thereby emit in the same color, the software cannot de-phase them or differentiate individual clusters[111]. In the original MiSeq and HiSeq series of instruments, Illumina uses 4 color chemistry for sequencing. Each nucleotide is labeled with a unique fluorophore that emits in a different channel (T – green, C – red, A – blue, G – yellow). In order to decrease sequencing time and costs, the next generation of instruments (NextSeq and NovaSeq) rely on 2 color chemistry (T – green, C – red, A – red and green, G – dark). This further impacts the ability of second generation instruments to differentiate homogeneous clusters, as is reflected in decreased Q-scores[112]. Updates in Illumina software, third-party algorithms, and modified experimental designs (such as inclusion of the high diversity PhiX spike-in) all improve the performance of low diversity libraries.

Experimental Considerations of scRNA-Seq

scRNA-seq experiments can be designed to answer diverse biological questions. Different approaches regularly obligate a trade-off between scale (number of cells profiled) and depth of coverage (number of genes detected per cell). For example, profiling the cellular composition of a tissue requires a large number of cells profiled at minimal depth, while monitoring cellular responses to stimuli requires fewer cells at a higher depth[113]. Optimally, the experimental design would allow for both large scale and high depth, however practically this is rarely possible[114]. Outlined below are technical considerations which are included in the design of scRNA-seq experiments.

Preparation of single cells: "Garbage in, garbage out"

Obtainability of single cells is a prerequisite to scRNA-seq. Most often, single cell isolation is accomplished by enzymatic digestion of tissue followed by a purification step (e.g. fluorescence activated cell sorting [FACS], density gradient centrifugation, or manual picking). Although often overlooked, this is a critical step that requires robust optimization. Key optimization points include: 1) minimization of cellular stress (low incubation temperature, short dissociation time, and gentle trituration), 2) maintenance of the transcriptional integrity of cells (inclusion of actinomycin D or RNase inhibitors), 3) optimization of the dissociation protocol for specific tissues and cell types (digestion conditions based on extracellular matrix composition, cell size and shape, and density of cell-cell contacts), 4) validation that live single cells are being used as input for the experiment (doublet discrimination, debris removal, live/dead cell stain, and FACS targeting). For cells obtained through LCMD, or any tissue which has been stored, fixed, or frozen, a test of RNA quality should be performed before the scRNA-seq experiment as degraded RNA can have a wide impact on the quality and applicability of data obtained[115].

Single Nuclei RNA-Seq

Individual nuclei sequencing (snRNA-seq) can replace whole cell sequencing in cases where healthy, intact cells are difficult to extract[116-119]. Specifically, nuclei isolations can be more robust

than cell isolations when working with frozen tissue (e.g. deteriorated cell membranes), highly myelinated tissue (e.g. spinal cord), or heterogeneous tissue containing delicate cell types (e.g. large adipocytes). In the central nervous system, snRNA-seq has been shown to be comparable to scRNA-seq in terms of transcriptional patterns of activation[120] and congruent with gene expression[121]. Naturally, the number of genes detected in snRNA-seq is less than in scRNA-seq, with more reads coming from intronic regions of pre-mRNA[121]. However, snRNA-seq experiments can be optimized to detect on average 4500 genes per nuclei (unpublished data), rivaling the depth obtained through some protocols of scRNA-seq. Furthermore, because of the high capture efficiency of pre-mRNAs, snRNA-seq is a strong platform for calculating RNA velocity. By comparing spliced and unspliced mRNA transcripts, RNA velocity can resolve the static transcriptional identity of a cell with the transcriptional trajectory it is poised to follow.[122]

Plate-based vs micro-isolation methods

Broadly, scRNA-seq experiments can be categorized into two pipelines: 1) monodispersed cells are FACS sorted into 96 or 384 well plates, with cDNA synthesis performed individually in each well (SMART-Seq[123], CEL-Seq[124, 125], SCRB-Seq[126], Div-Seq[116], Split-Seq[127]), and 2) monodispersed cells are loaded into a microfluidics device which either deposits cells into microwells or oil droplets, with cDNA synthesis performed individually in a microisolated environment (Drop-Seq[128, 129], Fluidigm C1[130], 10x Genomics[131]). Traditionally, plate-based approaches generate libraries with higher depth (~8,000 genes detected per cell compared with ~1000 using droplet methods), but are limited in the number of cells that can be processed simultaneously (96 or 384 compared with 100 - 100,000), and are more time consuming (several days compared to hours)[114]. Furthermore, commercial pipelines are sensitive to starting RNA content: they perform better with large cells that contain more RNA, and may struggle with small cells where RNA content is limited. snRNA-seq, for example, has not achieved high depth of coverage with droplet-based sequencing. Recently, a split-pooling method (Split-Seq[127]) has been established for generating libraries from over 100,000 individual cells in a plate-based format. This approach relies on paraformaldehyde fixation of cells to ensure that, instead of being

lysed, the permeablized cell membrane remains intact enough for cells to undergo three consecutive rounds of combinatorial indexing before library generation. The ultimate goal of all these approaches is to minimize cost per cell while maintaining library quality. Ultimately, the choice towards a particular method depends on the quantity of cells available, RNA content per cell, access to instruments, and depth of coverage required.

First strand cDNA synthesis: Reverse Transcription primers

Each cell contains a combination of information poor RNA (rRNA accounting for ~80% of total RNA, and tRNA for ~15%)[132] and information rich RNA (mRNA, miRNA, lncRNA, etc, combining to ~5% of total RNA). It is considered that an average mammalian cell contains 10pg of total RNA, and 0.1pg of mRNA[133]. While in bulk RNA-seq, rRNA is a useful measure of RNA quality (RIN) and is often a necessary step in determining how degraded the sample RNA is, for scRNA-seq it can compromise the information content in a library. Therefore, a rRNA depletion step or a mRNA enrichment step needs to be performed. For full transcriptome library preparation, mRNA enrichment is usually employed. To this end, the reverse transcription (RT) primer contains a 30 base pair (bp) 3' oligo-dT sequence which anneals onto the 3' poly-A tail of mRNA, but does not interact with rRNA or tRNA, effectively enriching for mRNAs via negative selection. In order to promote the primer binding at the start of the poly-A tail closest to the coding region, a V base (A, C, or G) is often included as the last base of the RT primer sequence. Upstream of this, the RT primer also contains a combination of: a unique molecular identifier (UMI)[134], a cell specific barcode (CSBC), Illumina adapter sequence, T7 promoter sequence, and a universal PCR primer sequence. The exact combination of these is determined by the protocol being employed.

UMI: A random 10bp sequence in the RT primer (NNNNNNNNNN). It allows a downstream analysis to quantify the exact number of original mRNA molecules that are represented in the amplified cDNA library, as each first strand cDNA molecule will incorporate exactly one UMI during the RT reaction. Given the large number of possible bp combinations (4^10, or 1,048,576) in an UMI, the chances that two mRNA molecules from the same gene will be tagged by an

identical UMI is minimal. Some protocols use shorter UMIs, arguing that even a 8bp UMI provides sufficient diversity for no two mRNA molecules from the same gene to be identically tagged.

CSBC: A sequence that tags all first strand cDNAs from the same cell with the same sequence. This allows cDNA from many cells to be multiplexed and processed together in downstream reactions, then assigned back to their cell of origin during data analysis. In plate-based methods, CSBCs are a known 6bp or 8bp sequence in the RT primer. Experimentally, each well of the plate receives a RT primer containing a distinct CSBC (for example, when working in a 96 well format, there will be 96 RT primers – one per well). The number of distinct RT primers determines the number of cells that can be pooled together in downstream reactions. In microfluidics-based methods, this is an undefined 12bp sequence in the RT primer; for each cell, it can be one of any 4^12, or 16,777,216, sequences. Probabilistically, this allows for hundreds of thousands of cells to be multiplexed together without CSBC duplication.

Illumina Adapter Sequence: For sequencing on an Illumina instrument, adapters complementary to those on the Illumina flow cell need to be present in the cDNA library. The 5' adapter is introduced to cDNA via the RT primer.

T7 Promoter: For those protocols that reply on linear T7 RNA polymerase amplification (CEL-Seq and CEL-Seq2), a minimal T7 binding sequence is introduced into the cDNA at the RT step.

Universal PCR Primer Sequence: For those protocols that reply on PCR amplification (SMART-Seq, Div-Seq, SCRB-Seq), a common PCR template sequence is introduced at both the 5' and 3' end of cDNA libraries. A single PCR primer allows amplification of cDNA with reduced primer dimer formation, improved sensitivity, and minimized spurious or nonspecific product formation[135]. In some protocols, the Illumina adapter sequence is used as the universal PCR template. The 5' PCR template is introduced into the cDNA via inclusion in the RT primer. The 3' PCR primer is introduced via a "strand switch" step[136-138]. Once the RT enzyme reaches the end of the RNA, its

inherent terminal nucleotidyl transferase activity nonspecifically extends the first strand cDNA by several nucleotides. Under proper conditions, this can be predominantly restricted to 3 cytosines[139]. A second primer, called the template switch oligo (TSO), contains 3 guanosines at its 3' end that allow it to anneal to the extended cDNA. Once the TSO is annealed, the first strand cDNA synthesis reaction can continue uninterrupted until the end of the TSO primer, thereby incorporating the reverse complement of the TSO template sequence into the 3' end.

cDNA amplification

Once barcoded cDNAs are generated, they can be pooled together and downstream amplification and library generation reactions can be considered "low input" as opposed to "single cell" (with the exception of Div-seq, in which cells are pooled post-tagmentation). This is an important distinction, as many enzymes are input-dependent and their efficacy is directly correlated with the concentration of starting material[140-142]. Only cells that have distinct CSBCs can be pooled, as redundant barcodes lead to convolution of downstream data analysis. Amplification is an important step in library preparations because sequencing instruments require a minimal library input concentration. The recommended amounts can be as low as 20nmol for one flow cell of an Illumina NextSeq 500/550 instrument, or as high as 930nmol for an S4 flow cell of an Illumina NovaSeq 6000 instrument (from Illumina Sequencing System Guides). In practice, this means that:

- for a pool of 100 cells that originally each contain 0.1pg of mRNA
- with 100% efficiency in conversion to double stranded cDNA
- counting 1 first strand cDNA molecule for 1 molecule of mRNA
- considering the average length of cDNA in a sequencing library is 500bp

the pre-amplified pooled cDNA – at 0.03236 fmol – would need to be amplified almost 1 billion times before it reaches the minimum recommended concentration for sequencing. PCR amplification requires 30 cycles to achieve this, giving the reactions plenty of opportunity to introduce bias into the final library (such as base pair mismatches, jackpot artifacts, and skewed

library representation)[143-145]. T7 amplification and split-pool qPCR have been implemented for improving the amplification step.

T7 Amplification: In place of ~10 cycles of PCR, commercially available T7 amplification kits are available to linearly amplify cDNA up to 1000 fold (NEB, Ambion). Because the T7 RNA polymerase produces aRNA (amplified, antisense RNA), these protocols require several additional steps which can increase the time to generate libraries by a full day: second strand synthesis, T7 DNA polymerase incubation, aRNA cleanup and fragmentation, ligation of a 3' adapter, and finally re-conversion to cDNA. Libraries generated with T7 amplification have been shown to have increased accuracy compared to PCR-based amplification protocols[146].

Split-pool qPCR: PCR reactions accumulate artifacts such as nonspecific template priming and chimera formation when cycled past the log-linear phase[147]. Given the amplification requirements for scRNA-seq library preps, large numbers of PCR cycles exceeding log-linear growth are often necessary. In order to avoid artifact formation, cDNA from a single library can be split between multiple PCR reactions and (with the addition of a dye such as EvaGreen) monitored on a real time thermocycler to ensure the reactions stay in logarithmic growth. Before the reactions start to plateau, they are pooled and redistributed as input to new PCR reactions. This can be repeated until the necessary library concentration is reached.

Library generation: fragmentation and indexing

Full length cDNAs (at ~2000bp) are too long to be sequenced on Illumina instruments. A fragmentation step, which can be in the form of a chemical or enzymatic reaction, is required for library generation. Proper fragmentation results in a library size distribution centered at ~500bp. At this stage, the 3' P7 Illumina adapter is inserted at the fragmentation site. The 5' P5 adapter from RT and the 3' P7 adapter are used for further rounds of library amplification via PCR. Both chemical and enzymatic fragmentation conditions can introduce unique biases into the libraries.

Chemical Fragmentation and P7 Adapter Insertion by Ligation: Chemical fragmentation, commonly achieved by incubation of RNA with divalent metal cations such as magnesium or zinc, is commonly employed in aRNA-based protocols[148]. Chemical fragmentation results in a largely random distribution of break points, which then serve as a blunt end for P7 adapter ligation. While the break points are random, RNA ligases are not fully efficient and introduce biases that propagate into the library[149, 150]. The ligated P7 adapter serves as the template for an RT primer, which converts fragmented aRNA into a cDNA library.

Enzymatic Fragmentation and P7 Adapter Insertion by Tagmentation: While a modified form of chemical fragmentation has been employed on cDNA[151], the most popular approach for fragmenting cDNA is tagmentation[142]. Tagmentation relies on a hyperactive variant of the Tn5 enzyme to simultaneously cleave DNA and insert an oligonucleotide containing the P7 sequence onto the broken ends[152]. Both commercial products (Illumina's Nextera and Nextera XT kits) and homemade recipes for transposome generation[142] are available. Tn5 is biased towards its preferred substrate, and can introduce insert/deletion errors in the cDNA; however, newly engineered Tn5 variants strive to overcome these problems[153, 154].

Indexing: The P5 and P7 adapter sequences on cDNA serve as a template for i5 and i7 Illumina indexes, which are introduced in primers during the final rounds of PCR amplification. Just like cells with unique CSBCs can be pooled, so can libraries with unique i5 and i7 indexes. Multiplexing libraries has the added benefit that it minimizes batch effects – technical variations during the sequencing run affect all pooled libraries evenly[155].

Ligation-Based Library Generation Methods

For small RNA library generation (miRNA, etc) in which templates do not have a poly-A tail, an alternative ligation-based approach can be used. This protocol relies on obstructing accessible rRNA sequences with a rRNA-specific blocking primer, followed by adapter ligation onto short RNA sequences[156]. Ligation is also the key component of first-strand cDNA generation in Split-

Seq, which has three sets of CSBC ligation primers. While most protocols rely on a cell membrane lysing and exposing RNA contents to the reaction buffer, Split-seq requires the cells to be fixed through paraformaldehyde fixation, with the cell membrane permeabilized but intact, effectively forming its own microisolated compartment with RNA immobilized inside. The first set of CSBCs, in this case better referred to as “well specific barcodes” (WSBCs), is introduced by the RT primer, similar to the aforementioned techniques except that thousands of single cells are present in each well and thus receive the same WSBC. Cells in all wells are then all pooled together, diluted, and randomly redistributed to new wells. Each cell still has an intact membrane and the cDNA generated in the previous step is hybridized to fixed RNA, preventing barcode swapping between cells. The second set of WSBCs is distributed and ligated immediately upstream of the first WSBC barcode. The cells are again pooled and redistributed into new wells, and the third WSBC is ligated immediately upstream of the second WSBC. The final product – cells in which the 5’ ends of cDNA have a consecutive sequence of 3 WSBCs – are then pooled and processed for library generation. Probabilistically, the chance of cDNA from two cells sharing the same combination of WSBCs is minimal.

Gene body coverage

Gene body coverage can be defined by the portion of a mRNA sequence that is recovered in a library. This is influenced by two factors: the size distribution of cDNA in a library, and a bias in the library generation. Size distribution is determined by the parameters of the sequencing instrument. For example, the optimal cDNA length for an Illumina sequencer is 400-600bp; this ensures efficient binding to a flow cell and bridge amplification. Most mRNAs are longer than this, and therefore must be sheared during library preparation. Given that the multiple sequences introduced via the RT primer are necessary for sequencing and data analysis, cDNA libraries generated via oligo-dT priming favor sequencing of 3’ (poly-A) ends of mRNA, leading to a 3’ bias in the gene body coverage. Information contained downstream of the shear site of cDNA is lost, because they do not contain the necessary information introduced by RT primers. Furthermore, because oligo-dT primers constitute ~40bp of non-biological sequence, on a paired-end

sequencing run these reads are often not mapped and instead used only for demultiplexing and analysis purposes. A 3' bias also affects mapping of splice sites, which are often located deeper into the transcript, and impinges on accurate detection of transcripts that have a conserved 3' end, such as protocadherins, because reads that have multiple alignments are discarded by mapping algorithms.

In order to circumvent this, libraries can be generated in a way that incorporates primer sequences into the shear sites, either through 1) CSBC-containing primer insertion via transposition (Div-seq), or 2) random hexamer, as opposed to oligo-dT, RT priming (Split-seq). Separate UMIs are not used with random primers, since absolute quantification of transcript number is impossible when an arbitrary number RT primers bind onto a single mRNA molecule. Random primers can also bind to rRNA molecules, and therefore a rRNA depletion step is often used to ensure libraries contain information-rich sequences. Commercially available methods for this include incubation with biotinylated primers antisense to rRNA sequences followed by a biotin-streptavidin pulldown (ThermoFisher Scientific's Ribominus, and Illumina's Ribozero), or addition of specific nucleases to digest away rRNA molecules[157].

Depth of Coverage

The depth of coverage in single cell libraries refers to the number of genes detected per cell. It is a function of library preparation (how many genes are present in the cDNA library) and sequencing (how many of those gene sequences are recovered). Increasing the amount of sequencing done on a cDNA library will improve recovery of genes until saturation is reached. At that point, effectively all the unique transcripts have been counted and further sequencing will only introduce duplicate reads that had been generated during library amplification. Saturation can be visualized by a saturation curve, which is generated by plotting the numbers of reads sequenced against the number of genes detected for each cell. Libraries sequenced to saturation provide the most rich biological information and introduce the least amount of artificial noise to downstream analysis.

The optimal amount of sequencing for a library is variable. High quality libraries (those consisting of thousands of genes) often require more sequencing to reach saturation than low quality libraries (those consisting of hundreds of genes). Likewise, in order to achieve the same number of reads per cell, libraries generated from thousands of multiplexed cells require more sequencing than libraries generated from hundreds of multiplexed cells. Traditionally, this has resulted in a tradeoff between sequencing depth and number of cells sequenced. Considering this balance, there are two schools of thought on how best to achieve biological significance through scRNA-seq: High coverage sequencing of fewer cells vs low coverage sequencing of more cells. Both have their merits, and the decision to proceed with one approach or the other largely depends on the biology in question. Generally, low depth of coverage is sufficient for characterizing cellular diversity in a population where large numbers of cells needed and cell markers are abundantly expressed, while high depth of coverage is required for profiling heterogeneity in a subpopulation of similar cells or catching transcriptional changes in genes that have low expression.

Technical Considerations for scRNA-Seq

Over 20 experimental approaches to scRNA-seq are currently available. Naturally, all of them rely on maintaining a clean environment, such as a dedicated hood, to provide protection from exogenous RNase and DNAse contamination. Minimizing handling time and working at low temperature are also both critical to generating high quality data. Given these preconditions, there are a number of technical adjustments that improve on the performance of published protocols. The two most common of these are reaction conditions and primer modifications.

Reaction conditions

Enzyme choice, specifically for RT, ligation, IVT, and PCR reactions, varies between protocols. Hundreds of enzymes are commercially available, but only a handful are used for single cell purposes. The selection criteria for these enzymes include fidelity, efficiency, thermostability, processivity, speed, and RNase and template switch activity. As new and improved enzymes are

developed, they are employed in successive versions of existing protocols. Consistency is key after optimizing a protocol for an enzyme of choice, because different enzymes can introduce distinctive biases into data.

Substrate concentration is a fundamental factor that affects reaction efficiency. Crowding agents, such as PEG, can increase the effective substrate concentration, and have been implemented in some scRNA-seq protocols. Equally effective, and more cost-efficient, is decreasing the reaction volumes. Smaller reaction volumes can only be applied when RNA or cDNA is the limiting reagent, and evaporation volume should be accounted for thoroughly as it can represent a large proportion of the reaction volume. There are a number of compounds that have been shown to improve PCR efficiency, especially for GC rich regions, by decreasing the melting temperature of DNA. Among them are DMSO and betaine, the first of which is commonly included in GR-rich PCR reaction buffers, and the second of which has been used as an additive in some scRNA-seq protocols. Unexpectedly, however, betaine has been shown to inhibit PCR efficiency in some cases and should be used with care[158]. Likewise, the concentration of $MgCl_2$ can be adjusted in PCR reactions to improve efficiency, but high concentrations will act as an antagonist.

Loss of transcripts after successful cDNA synthesis also decreases depth of coverage, and largely occurs during purification steps which rely on a minimum input concentration to be effective. Two purification strategies – columns and SPRI beads – are most commonly used in scRNA-seq experiments. RNA and DNA binding columns are especially sensitive to high volumes and low concentrations, and even low-input grade commercial columns will lose transcripts in the flowthrough. Therefore, SPRI beads have become the routine purification method for RNA and DNA. The majority of transcript loss from SPRI bead purification occurs when they are used for size selection, as lower ratios of SPRI buffer to reaction volume result in higher molecular weight capture cutoffs. This property allows them to be used for primer dimer removal, but also removes lower molecular weight library. SPRI beads optimized for RNA or DNA cleanup are commercially available, but can also be made in house for a fraction of the cost.

Primer Modifications

Apart from sequencing, the most expensive part of scRNA-seq is the initial primer cost. All RT primers require HPLC purification to ensure they are RNase free, and many primers incorporate modifications or non-standard bases such as iso-bases and locked nucleic acids (LNAs). Furthermore, each protocol uses a unique combination of primers that are not interchangeable. Primers are designed to be as short as possible because, especially for the TSO, length is inversely proportional to capture efficiency[159]. Primer modifications can be broken into two groups: those that enhance efficiency, and those that minimize the potential of primer concatamerization.

Two common modifications are used to enhance efficiency. First is inclusion of LNA nucleotides in the TSO primer. LNAs are synthetic nucleotides with stronger hybridization properties than traditional bases[160]. Oligonucleotides containing LNA bases therefore have a significantly higher melting temperature, and LNA guanines (+G) have been included in the 3' end of TSOs to allow the short (rGrG+G) sequence to efficiently bind to the cytosines (dCTPs) introduced to first strand cDNA via nucleotidyl transferase. Alternatively, all three 3' guanine bases can be ribonucleotides (rGrGrG), which also form tight bonds with dCTPs. The second modification is inclusion of phosphorothioate bonds between the five 3' bases of the PCR primer. These bonds make the oligonucleotide resistant to exonuclease degradation. An Exo1 incubation is commonly included post-RT to degrade free primers, and carryover of Exo1 would degrade any subsequently added PCR primer lacking this modification.

Exo1 treatment is an effective way to degrade free single stranded primers, thereby preventing downstream dimerization. Primer dimers outcompete cDNA in PCR reactions, making library amplification difficult. Exo1 treatment does not, however, prevent primer concatamerization in which oligonucleotides serially prime off each other to form a repeating double stranded sequence. Primer concatamers appear as a hedgehog pattern on a bioanalyzer trace and, due to

both their wide size distribution and preferential amplification, are a source of library contamination that is impossible to remove. Two primer modification strategies are commonly used to minimize primer concatamerization. First is blocking the 5' end of a primer through attachment of a protein such as biotin, which physically impedes interactions with another oligonucleotide. Second is incorporation of isoguanine (iso-dG) and isocytosine (iso-dC) bases at the 3' end. These bases can only form hydrogen bonds with one another, and therefore cannot serve as a template for extension.

Computational Considerations of scRNA-Seq

During a sequencing run, the data is in bcl format: a binary file which contains raw data from the detector's measurement of fluorescence. This is converted for end users into FASTQ format, a text-based file which contains the sequence information for each read along with the "Q score", or quality score, for each base. The FASTQ file then needs to be demultiplexed by indexes and CSBCs, aligned to the genome, assigned to a transcript, and counted. This process can be very computationally expensive and time consuming, and a large number of algorithms are being developed and improved to streamline the experience.

Q score and read trimming

The Q score (or Phred quality score) reflects the probability with which a base is called correctly by the sequencer. A Q score of 30 represents a 1/1000 chance that the base was incorrectly called, aka 99.9% accuracy. This is the standard cutoff for processing reads. Those reads that contain bases with Q <30 can be discarded from analysis. The quality of reads tends to drop over time, and bases deeper into read 2 have a lower Q score than the starting bases or bases in read 1[161]. Therefore, long read 2 sequences (>50bp) are often trimmed with a Q score threshold before analysis to retain the most accurate information.

Demultiplexing

Illumina automatically separates reads based on Illumina indexes, however CSBCs within each index need to be demultiplexed by the user. For paired end runs, both the reads (only one of which contains the CSBC) need to be demultiplexed. There are a number of scripts available on GitHub to do this[162, 163]. When CSBCs have a known sequence, two pieces of information should be in place for the algorithms: the number of expected barcodes or the sequences of the barcodes, and the hamming distance between barcodes. The hamming distance is the number of mismatches between barcodes. Practically, setting a hamming distance of 2 allows a script to assign a read to a barcode even if it has 2 mismatches to the known barcode. This is only acceptable when the hamming distance between the barcodes is 4 or more, as the script could otherwise erroneously assign a read to the wrong barcode. When CSBCs have an unknown sequence, other programs can be employed for demultiplexing (FastqMultx, Picard Tools, FASTX Toolkit, UMI tools). Most of these programs are multipurpose and also include scripts for read trimming, gene counting, calculating gene body distribution, and other basic analyses.

Sequence Alignment

Aligning read sequences to a reference genome or transcriptome, aka data mapping, requires a complex algorithm that takes into account mismatches, splice junctions, intronic sequences, and read directionality. Traditional algorithms such as Bowtie[164] and BWA were based on Burrows-Wheeler transformation due to its computational efficiency[165]. With increased memory in computers, newer approaches such as STAR (Spliced Transcripts Alignment to a Reference)[166] can accelerate the time it takes to map data from hours to minutes. STAR sequentially searches through an uncompressed suffix array for maximally mapping sequences, and then clusters and stitches them together. Alternatively, pseudo-alignment methods such as Kallisto infer gene identity by checking the similarity of short, known sequences with transcripts and directly outputting a counts file[167]. Pseudo-alignment algorithms have the benefit of needing little computational space, and can be run on laptop computers as opposed to multi-core workstations. There are two commonly used reference genome annotations: Ensembl[168] and gene symbol. In Ensembl, each transcript variant (eg alternatively spliced variant) is given a unique transcript ID

and a common gene ID. With gene symbol, all the transcript variants for a single gene are assigned to the same gene symbol. The appropriate reference annotation should be determined prior to mapping, as separate alignments for splicing information can either convolute or be necessary for data analysis (e.g. differential expression), depending on the model. Reads that do not align to a reference genome are usually due to primer concatamerization. Thus, the percent alignment is a strong indicator of how well the experiment was designed and executed. A low percent alignment (<70%) indicates that there is room for optimization in the experiment. The raw number of aligned reads per CSBC is also a computational verification that single cells were used as input. An outlier population of CSBCs that map to an excessive number of reads can indicate that doublets were processed in the experiment.

Gene Counting

Once sequences have been mapped, the number of reads that aligned to each gene can be counted and a spreadsheet can be generated with gene expression information. This spreadsheet containing raw gene counts for each cell is called a counts matrix, and is the input for downstream data analysis. The absolute number of mRNA molecules present and processed at the RT step can be approximated by reducing the counts matrix by UMIs. UMI correction is useful in minimizing jackpot artifacts, in which by random chance a transcript is amplified over and over during PCR[162]. Commonly, the number of genes detected per cell is determined at this step. This number can be inflated in two ways, first by considering Ensembl transcript variants as separate genes, and second by considering UMIs detected instead of genes detected. A saturation curve can also be generated at this step, and is useful for determining whether or not additional sequencing runs would lead to greater biological insight (in the form of more genes detected per cell). Transcript drop out, or false negative expression, is the most significant source of noise in single cell experiments. It is especially prevalent in low coverage data sets, and obscures meaningful data that could be accessed with higher coverage[169]. The only definite way of minimizing drop out effects is by increasing the number of genes detected per cell, either by experimental design or sequencing strategy.

Filtering and Normalization

Gene counts can be used raw, but aside from UMI correction there are a number of normalization techniques that should be considered for accurate analysis. Transcripts per million (TPM) accounts for the total number of reads detected in a cell, and normalizes expression values based on the proportional representation of each transcript in each cell. TPM can artificially inflate the calculated expression of genes in cells that have few genes detected overall, so a filtering step should be implemented to discard outlier cells with few genes detected. Reads per kilobase per million mapped reads (RPKM) is similar to TPM, but also takes into account the length of the transcript. RPKM is useful for normalizing reads that come from multiple sites on a single transcript, such as with random hexamer priming or full cDNA tagmentation, and should only be utilized with experimental protocols that do not rely on digital 3' counting. ERCC spike-in normalization calculates the relative expression of detected genes to a series of known RNA spike-ins for each cell. ERCC normalization accounts for variability in the efficiency of RT reactions for each cell, and relies on the measured concentration of ERCCs falling along a linear curve to calculate the expression level of genes. Finally, downsampling is used when there is a disparity in the number of reads mapped between cells. Downsampling computationally imitates lower-depth sequencing conditions and results in a user-determined number of reads mapped per cell. It can both help normalize gene expression and mitigate the effects of dropout by equalizing dropout effects between cells.

Traditional normalization methods that rely on housekeeping genes (such as actin or GAPDH) are not feasible for scRNA-seq data. At the single cell level, mRNA synthesis occurs in transcriptional bursts, which result in huge stochastic variability of the number of mRNA molecules for any given transcriptional locus present between cells[70]. However, housekeeping genes do play a role in scRNA-seq analysis. Lack of expression of a set of housekeeping genes, including cell type specific genes that are known to be abundantly present (e.g. Malat1 for neurons), is useful for filtering out reads that come from non-cellular origins, such as processes or other debris. Furthermore, the transcriptional profiles of cells that were damaged during

dissociation are dominated by the expression of stress markers such as Il1a, and should also be filtered out to avoid confounding downstream data analysis.

scRNA-Seq Data Analysis

While the analyses associated with scRNA-seq and bulk RNA-seq experiments share similarities, protocols cannot be reused between the two. Similarly, computational approaches designed for bulk RNA-seq data analysis cannot be applied to scRNA-seq datasets because single cell transcriptional profiles are noisy, sparse, and contain more sequencing artifacts than bulk data. In order to find biological information in scRNA-seq data, over a hundred algorithms have been developed to separate signal from noise and dropout, identify expression patterns across samples, and delineate transcriptional pathway activation. Many of these approaches are repurposed from pre-existing algorithms for mathematical models for non-biological data. Broadly, statistical analysis of scRNA-seq datasets depends on two factors: similarity matrix generation and dimensional reduction. Similarity matrixes can be built off correlation or distance metrics. Dimensional reduction can be a linear or non-linear function, and is applied to project high dimensional data onto a visualizable two dimensional graph. This visual representation takes the form of either cell clustering or cell ordering along a trajectory.

Similarity Matrixes

A scRNA-seq dataset is a large matrix, with the dimensions being number of cells profiled x number of genes detected. In high dimensional space (with each gene representing its own axis), a cell can be considered a point defined by the coordinates of its gene expression. The purpose of a similarity matrix is to determine how closely cells are related to one another in the high dimensional space. This can be accomplished through distance metrics (eg, Euclidean distance), which calculates cell-cell similarity based on the number of transcripts counted for each gene, or through correlation metrics (eg, Pearson's correlation), which considers the relative expression of each gene. Given the logarithmic amplification schemes (PCR) used for generating scRNA-seq data and the non-linear relationships between transcriptional expression an biological function,

correlation metrics have proven more reliable than distance metrics for generating similarity matrixes[170].

Pearson's correlation is currently the most common non-linear metric employed by scRNA-seq algorithms and works well on heterogeneous samples such as those generated for cell atlas projects, which aim to identify discrete populations of cell types. Other approaches have been implemented for projects that have a more homologous starting population. Density and centrality metrics (eg, L-infinity[171, 172]) and mutual information (eg, Jaccard[173]) have revealed important biological information from highly similar cellular populations. Mutual information is especially useful for parsing information from a relatively homogenous population of cells, such as transcriptional alterations occurring in a single cell type.

Dimensional Reduction

Dimensional reduction algorithms determine a subset of genes that account for the greatest amount of variance in the overall dataset, and use these genes to project the high dimensional point cloud data onto a two-dimensional graphic representation. The goal is to retain as much information about the full dataset as possible while considering as few variables as possible. Principal Component Analysis[174] (PCA) is the most commonly used linear dimensional reduction algorithm. In PCA, highly expressed and variable genes are binned into principle component (PC) groups that explain variation along a single axis. The first PC often dominates the analysis, and may contain genes that are not biologically relevant to the study. Cell cycle genes and stress markers arising from poor dissociations are examples of this[175]. Such genes can be excluded by pre-filtering the data, or excluding the first PC from downstream analysis. The statistically significant PCs, those that should be included in downstream analysis, can be determined through a jack-straw plot or an elbow plot (REF). Both of these charts show the extent to which each PC accounts for variability in the data. PCA itself is not a visualization tool, but rather its output can be used as input for visualization algorithms. Multidimensional Scaling[176] (MDS) is a

non-linear dimensional reduction tool similar to PCA, except its focus is to retain similarities in the data while PCA aims to retain variability.

T-distributed Stochastic Neighbor Embedding (t-SNE) is another commonly used non-linear machine learning approach for dimensional reduction and data clustering[177]. t-SNE performs pairwise comparisons of cells in high dimensional space and uses different transformations on subsets of cells to accurately project their relationships onto two dimensions. It relies on a user defined value, perplexity, to identify how many neighbors a cell has, and from there to define clusters. Because t-SNE requires user input, the output representations can be misleading; however with appropriate care, t-SNE is a valuable algorithm for scRNA-seq analysis. In essence, t-SNE generates cell groupings in a dimensionally reduced representation and calculates the error between this grouping and that found within the high dimensional space. Groupings are iteratively adjusted in a directed manner to minimize this error, known as the Kullback Leibler divergence.

Supervised dimensional reduction is also possible when the data contain known and identifiable elements. Linear Discriminant Analysis (LDA), closely related to PCA, builds models on explicitly defined data, and has been used as both a filtering algorithm and a tool to merge cell clusters from concatenated datasets[178, 179]. Alternately, dimensional reduction algorithms can be confined to consider a list of pre-defined genes as input. The results of artificially restricted analyses are always biased and suffer from loss of information, but may be useful in answering targeted questions.

Two Dimensional Representations: Visualization and Analysis

Once dimensional reduction has been executed, data can be plotted, visualized, and analyzed. The two most common visualization techniques are clustering and ordering. Clustering can be done either through scatterplots, where each cell is represented by a point on a graph, or through heatmaps, where each cell defines a row or column. The x- and y- axes of scatterplots are effectively arbitrary units and do not serve any measurement purpose. Instead, they provide a

reference point for observing the location and organization of cells on the graph, which reflect the relationships between them. The physical distance between clusters on a scatterplot also does not have intrinsic meaning, but rather is useful to clearly delineate clusters. In heatmaps, cells are hierarchically clustered based on transcriptional profiles such that transcriptionally similar cells are adjacent to one another. The heatmap may be a correlation matrix, in which case the rows and columns are symmetrically organized cells and the colors represent the global correlation between pairs of cells. A heatmap is also useful to show the expression of multiple genes in the dataset simultaneously. In this case, the heatmap color indicates the expression level of a gene (organized in rows) for each cell (organized in columns).

Differential expression analysis between clusters allows identification of marker genes for each cluster, thereby revealing the identity of cells in the cluster[128, 180]. It also allows the determination of gene coexpression between clusters. This approach is well suited for studies that aim to classify cell types in a heterogeneous population. However, projects that aim to study transcriptional transitions, for example during development or disease progression, face two innate setbacks in clustering algorithms: 1) they break some transcriptional relationships in favor of others, and 2) they struggle with inherent continuity within data. Cell ordering algorithms have been developed to overcome this[181, 182]. Instead of breaking cells into discrete clusters defined by marker genes, ordering algorithms search for continuous and overlapping patterns of expression, calculating a backbone that traces the patterns, and aligns cells along this path to determine a transcriptional trajectory. Cell ordering algorithms can struggle in building the backbone when it involves a bifurcation or loop, and specialized algorithms have been developed to trace processes that are known to split into two or more populations[183-185]. However, these algorithms can force the data into bifurcations that may not be biologically accurate. Topological data analysis (TDA[186]) is a branch of applied mathematics that creates two-dimensional representations of high dimensional data in a way that maintains not only pairwise relationships, but also the overall structure of the data. Uniquely, TDA applied to scRNA-seq datasets (scTDA[74]) is able to recapitulate the shape and trajectories of biological processes without any pre-defined assumptions or inputs.

Alternative representations of single cell data may be generated using k-Nearest Neighbor Graphs (k-NNG) and Force Directed Graphs. k-NNGs constitute a supervised machine learning algorithm used for both classification and regression. Here a pre-specified set of K groups are assumed. In the instance of single cell sequencing data, k-NNGs are utilized to classify individual cells into similar groups. The key assumption underlying k-NNGs is that highly similar cells with reside near one another in a low dimensional representation. Proximity is therefore related to similarity. In the k-NN algorithm, data is loaded, then K is initialized to specify a chosen number of neighbors. Subsequently, for each example cell within the sample, the distance between the query cell and the example cell drawn from the data. An ordered collection is created, consisting of the distance and index of the query and currently evaluated example from the data. The ordered collection is sorted by distance and index. The first K entries from the sorted collection are then taken and the labels of individual cells are recorded. k-NNGs are particularly suited to noisy and large datasets, but inherent disadvantages arise owing to the need to specify k and the computational cost associated with classification. Force directed graphs provide a means of generating k-NNGs in a format that is considered aesthetically pleasing and accounts for topological features present within continuous single cell data. Such graphs enable the recapitulation of deterministic long distance relationships between cells that are often obscured in t-SNE plots. However, force directed layouts are computationally intensive (often limited to datasets of 10,000 cells), diminishing their scalable utility for large scale studies.

Conclusions

Single cell RNA-sequencing is an exciting direction in science that has opened the doors to many important discoveries. As the field evolves, new experimental and computational approaches will make it possible to answer more outstanding questions in biology. Progress in being made on several frontiers, pushing the limits of what is currently technically possible. Spatially resolved sequencing, which allows simultaneous measurements of cellular orientation in tissue and

transcriptional profiling, is underway[187, 188]. Spatially resolved sequencing can provide direct information on intercellular interactions as well as intracellular changes. Combined scRNA-seq and single cell chromatin accessibility, or computational algorithms that link the two modalities, allows for interrogation of cell-specific mechanistic regulation[189, 190].

Chapter 2: Applied Topology Reveals Developmental Progression of mESCs With Single-Cell Resolution

Introduction

The differentiation of motor neurons from neuroepithelial cells in the vertebrate embryonic spinal cord provides a well characterized example of cellular lineage commitment and terminal cellular differentiation[2]. Neural precursor cells differentiate in response to spatiotemporally regulated morphogen gradients generated in the neural tube by activating a cascade of specific transcriptional programs[2]. A detailed understanding of this process has been limited by the inability to isolate and purify sufficient quantities of synchronized cellular subpopulations from the developing murine spinal cord. Alternatively, *in vitro* approaches have provided the opportunity to study the basic mechanisms of motor neuron differentiation at a cellular level[34], and to gain insights into motor neuron disease mechanisms[40, 65]. An inherent limitation of the *in vitro* approach is the differential exposure of embryoid bodies (EBs) to inductive ligands and uncharacterized paracrine signaling within EBs, which lead to the generation of heterogeneous populations of differentiated cell types[191]. Consequently, motor neuron disease mechanisms must be studied in a heterogenous background of cell types whose contributions to pathogenesis are unknown. The ability to interrogate the transcriptome of individual differentiating cells in this context could provide fundamental insights into the molecular basis of neurogenesis and motor neuron disease mechanisms.

Single-cell RNA-sequencing conducted as a function of time affords the opportunity to dissect complex transcriptional programs and their impact on cellular differentiation of individual cells, capturing heterogeneous cellular responses to developmental induction. Several algorithms for the analysis of single-cell RNA-sequencing data from developmental processes have recently been published[181, 183-185, 192, 193]. These methods can be used to order cells according to their expression profiles, and make possible the identification of lineage branching events. However, some of these approaches lack an unsupervised framework for determining the transcriptional

events that are statistically associated with each stage of the differentiation process[192]. With some methods, the statistical framework is strongly biased, for example by assuming a differentiation process with exactly one branch event[183, 185] or a tree-like structure[181, 193]. This is problematic in settings where the biological process does not fit the assumed model or is encumbered by noise. Moreover, the vast majority of these methods do not exploit the temporal information available in longitudinal single cell RNA-sequencing experiments, and require the user to explicitly specify the least differentiated state[181, 183-185, 192]. Here we report an unbiased, statistically robust mathematical approach to single cell RNA-sequencing data analysis that addresses these limitations.

Topological data analysis (TDA) is a nascent branch of mathematics directed towards studies of the continuous structure of high-dimensional data sets. TDA has been utilized to study viral re-assortment[194], human recombination[195, 196], cancer[197], and other complex genetic diseases[198]. We build upon TDA, and introduce a method (scTDA) that enables the unbiased study of time-dependent gene expression using longitudinal single-cell RNA-seq data. The scTDA method constitutes an improvement over previous approaches in that it provides a robust, unsupervised, statistical framework for the detection of transient cellular populations and their transcriptional repertoire, without assuming a tree-like structure for the expression space or a specific number of branching points. scTDA provides all the necessary components to assess the significance of topological features of the expression space, such as loops or holes. In addition, it exploits chronological experimental information when available, inferring from the data the least differentiated state.

Using scTDA, we dissect the specific transcriptional programs that regulate developmental decisions as mESCs transition from pluripotency to fully differentiated motor neurons and concomitant cell types. We comprehensively characterize the dynamic appearance of mRNAs encoding signaling proteins, transcription factors, RNA splicing factors and long non-coding RNAs (lncRNAs) during the transition from pluripotent cells to neural precursors, progenitors, motor

neurons, and interneurons, thus providing a valuable resource for studies of stem cell differentiation and neurogenesis, and more generally to any cellular differentiation process amenable to single cell transcriptomic analysis.

Results

Analysis of Continuous and Asynchronous Processes Using scTDA

Single-cell expression can be represented as a sparse high-dimensional point cloud, with the number of dimensions being equivalent to the number of expressed genes (~10,000). Extracting biological information from these data requires reducing the dimensionality of the space. Widely-used algorithms, such as multidimensional scaling (MDS), independent component analysis (ICA), and t-distributed stochastic neighbor embedding (t-SNE), have been adapted to flow cytometry, mass spectrometry[193, 199], and single-cell RNA-seq measurements[181, 200]. These strategies, however, were not designed to preserve continuous relationships in high dimensions. Figure 2a represents a simple example of a one-dimensional continuous manifold (a circle) twisted in three dimensions. Reduction to two dimensions using these algorithms introduces artifacts in the low-dimensional representation, including artifactual intersections (in MDS and ICA), and tearing of the original continuous structure apart (in t-SNE). As cell differentiation is a continuous asynchronous process, we reasoned that longitudinal single-cell data would be best analyzed using a mathematical approach that accounts for continuous structures.

We developed a computational approach to the analysis of longitudinal single-cell RNA-seq data based on the TDA algorithm Mapper[201] (Supplementary Note 1). Mapper builds upon any given dimensional reduction algorithm, such as MDS, and produces a low-dimensional topological representation of the data that preserves locality. To that end, the projection obtained by the dimensional reduction algorithm is covered with overlapping bins, and clustering of the data within each bin is performed in the original high-dimensional space (Figure 2b). A network is constructed by assigning a node to each cluster, and clusters that share one or more cells are connected by an edge. The result is a low-dimensional network representation of the data where

nodes represent sets of cells with similar global transcriptional profiles, and edges connect nodes that have at least one cell in common (Figure 2b). An important feature of these networks is that elements that are connected in the low-dimensional representation lie near each other in the original high-dimensional expression space, contrary to what occurs with most of the dimensional reduction algorithms currently in use (Figure 2a). This construction is thus robust against complex structures in the high-dimensional expression space, including non-tree-like trajectories (Figure 2a). Additionally, since nodes represent clusters of cells, the approach is highly scalable to large datasets.

Based on these considerations, we adapted the use of topological representations to the analysis of longitudinal single-cell RNA-seq data, introducing the necessary mathematical concepts and statistics (Supplementary Fig. 1). To identify expression programs associated with different parts of the topological representation without predefining any cellular population, we developed the concept of gene connectivity. The connectivity of a gene acquires a significant value when cells that share similar global transcriptional profiles express that particular gene more than random (Figure 2c). Genes with a significant connectivity are expected to be distinctively involved in a particular stage or stages of the differentiation. To identify the least differentiated state in the topological representation, we introduced the notion of root node, as the node that maximizes correlation between sampling time and graph distance. Using this root node as a reference, we computed the centroid (Figure 2d) and dispersion (the standard deviation) of genes in the topological representation. Genes with low centroids are upregulated in stem-like cells, whereas genes with large centroids are specific to fully differentiated cells. To identify transient cellular subpopulations arising throughout the differentiation, we clustered low-dispersion genes with a significant connectivity in the topological representation based on their centroid (Davies-Bouldin criterion, Figure 2e). Finally, to assess the significance of topological features of the representation, such as loops and holes, we used persistent homology, a framework within TDA for deriving and classifying topological features associated to data (Supplementary Note 1). Further details of the scTDA method and its mathematical foundations can be found in Supplementary Note 1.

scTDA Orders Asynchronously Differentiating Cells in Time

We assessed the capacity of scTDA to order asynchronously differentiating cells using a controlled setting where the truth is known. To that end, we first simulated a noisy, branched, asynchronous cellular differentiation process with two branching points (Figure 3a). 700 cells were sampled at three time points and the expression levels of 500 genes were simulated. From this data, scTDA correctly reconstructed the topological structure of the differentiation tree, and identified the most stem-like state (Figure 3b). In contrast, the algorithms Diffusion Pseudotime[183], Wishbone[185] and Slicer[184] failed to correctly assign some of the branches, producing artifacts in the inferred pseudo-temporal ordering of the cells (Figure 3c to 3e), even if the most stem-like state was provided by the user of these algorithms.

To quantify the performance of scTDA more comprehensively, we extended the above simulations to processes with one, two, or three branching points. scTDA showed higher correlation between the inferred pseudo-time and the actual simulated differentiation time, reconstructing the underlying differentiation process more accurately from the data (Figure 3f). The improvement was particularly large when drop-out events were included in the simulation (Figure 3g), where the graphical representation produced by Diffusion Pseudotime, Slicer, and Wishbone were often unable to correctly infer the structure of the differentiation tree (Supplementary Fig. 2).

Single-Cell RNA-Sequencing during Motor Neuron Differentiation

We applied our method to experimental data collected from longitudinal single-cell RNA-seq of *in vitro* motor neuron differentiation. We differentiated mESCs into motor neurons using a well-established protocol[34]. The mESC line expresses enhanced green fluorescent protein (eGFP) under the control of the early motor neuron-specific promoter *Mnx1*, allowing identification of committed motor neurons. Using modified CEL-Seq approach[124], we generated cDNA libraries from 2,744 single cells, including two biological (non-technical) replicates, sampled across days two through six of differentiation, spanning the transition from pluripotency to post-mitotic commitment. The differentiating cells were concomitantly sampled in bulk at each time point. To

assess optimal depth of sequencing coverage, reproducibility, and venues for experimental improvement, we initiated our study with a pilot experiment using a modified CEL-Seq protocol over 440 cells sampled throughout the time course of motor neuron differentiation. We sought to improve amplification and increase statistical power by sequencing more single cells per time point with the depth of sequencing coverage near the saturation point. To accomplish this, we modified the CEL-Seq protocol and sequenced 2,304 individual cells in a separate differentiation. In this study, herein referred to as the main experiment, we detected 19,009 transcripts across the dataset. Given the improved statistical power associated with the main experiment, we focused our attention on the results from the main study, and assessed the reproducibility of the results using the pilot study.

Identification of Cell Populations and Surface Markers as a Function of Differentiation

We used scTDA to analyze the two longitudinal single-cell motor neuron differentiation RNA-seq datasets. We filtered the data, retaining, respectively, 373 and 1,964 individual cells that passed stringent quality control tests (Supplementary Fig. 3). The scTDA analysis recapitulated chronological order based on mRNA expression levels alone (Figure 4a and 4b, and Supplementary Figs. 4 and 5), and simultaneously detected detailed transcriptional relationships between individual cells. We observed only a mild dependence of the library complexity on differentiation time (Supplementary Fig. 6), unlike other experimental settings[202, 203]. We did not observe the presence of large batch effects (Supplementary Fig. 7). Compared to representations of the same data generated by PCA, MDS, t-SNE, Monocle, Wishbone, Slicer, and Diffusion Pseudotime, we found that scTDA best preserved the continuous chronological structure of the differentiation process (Figure 4c and Supplementary Figs. 8 and 9).

Our analysis identified 7,620 genes with significant gene connectivity (q-value < 0.05, Supplementary Fig. 10), compromising 74% overlap with the pilot experiment (p-value < 10^{-100}, Fisher exact test, Figure 4d). This large number of significant genes is indicative of the transcriptional heterogeneity of the dataset, encompassing a large spectrum of developmental stages. The centroids of these significant genes were consistent across the two experiments

(Pearson correlation = 0.9, *p*-value < 10^{-100}, Figure 4d), providing another consistency check of our approach. The validity of our approach was also confirmed by the progressive expression of known markers associated with pluripotent cells, progenitors and mature motor neurons over time (Figure 4e and Supplementary Fig. 4c), as well as downregulation of cell cycle genes in the post-mitotic population of neurons (Supplementary Fig. 11). These results were stable under different choices of parameters (Supplementary Fig. 12).

Based on the distribution of the centroid and dispersion of genes (Supplementary Fig. 13), scTDA revealed four transcriptionally distinct cellular populations arising throughout the differentiation (Figure 5a). In total, 488 genes were assigned to four principal expression groups corresponding to these cellular populations (Figure 5b). Groups 1a and 1b contain genes uniquely expressed within early EBs (*Oct3/4*$^+$ cells), corresponding to pluripotent and neural precursor states. Genes in groups 2 and 3 are uniquely expressed within the progenitor (*Olig2*$^+$ cells) and post-mitotic ensembles (*VAChT*$^+$ cells), respectively. Ontology enrichment analysis showed an enrichment for developmental genes and genes related to DNA replication in groups 1a, 1b, and 2, whereas group 3 is enriched for genes related to axonogenesis, neuron migration, and regionalization, consistent with the underlying cellular differentiation process.

Our analysis revealed several transcripts that encode proteins with an extracellular domain, which constitute suitable cell surface marker candidates for sourcing niche populations for further study. Several of these surface markers had not been previously reported in this biological context. We validated the presence of these markers *in vitro* and *in vivo*, using immunohistochemistry in EBs (Figure 5c, and Supplementary Fig. 14) and in the murine embryonic spinal cord (Figure 5d).

Topological Characterization of Distinct Proliferative States

A remarkable feature of the topological representation of the motor neuron differentiation data is the presence of numerous loops in the representation of the neural precursor population (Figure 5a), in contrast to the mESCs and motor neurons. This feature was also observed in the

topological representation of the pilot experiment (Supplementary Fig. 4). Neural precursors are rapidly proliferating as a consequence of the induction with retinoic acid (RA). To evaluate whether the loops in the representation of neural precursors are caused by differences in the cell cycle, we built a new topological representation using only cell cycle genes (Supplementary Fig. 15a). Consistent with this hypothesis, the new topological representation contained substantially larger loops in the same region (Supplementary Fig. 15a), separating *Stra8*^{up} neural precursors and progenitors into proliferative and non-proliferative populations according to the expression of *Mki67* and other markers of proliferation (Supplementary Figs. 15b and 15c). We used persistent homology to assess statistically the significance of these loops in the topological representation, asking whether loops of similar size could arise from noise effects. This analysis showed a strong statistical significance for some of the larger loops (q-value < 0.05, Supplementary Fig. 15d), consistent with a biological, rather than technical, origin for these features.

Characterization of Developmental Transitions

We next used scTDA to characterize the developmental transitions that occur during the differentiation of motor neurons. These results constitute a significant resource for studies of stem cell differentiation and neurogenesis, and we have made them available through an online database. We summarize here some of the main transcriptional programs that are associated to the transitions between the four identified transient cellular populations (Figure 6a).

Expectedly, the transition from a pluripotent to a neural precursor population is characterized by the transcriptional dynamics of pathways involved in RA signaling and downstream effector proteins[204, 205] (Figure 6a). Our analysis resolved with unprecedented resolution the timing of the transcriptional events occurring during this transition, identifying upregulation of *Stra8* and downregulation of *Fgf4* as some of the earliest events that mark the transition (Supplementary Fig. 16). Subsequently, there is transcriptional activation of a subset of the homeobox gene family, including *Hoxa1* and *Hoxb2-8*, which continue to be expressed during later stages of the

differentiation process, and *Hoxb1*, which is transiently expressed along with caudalizing transcription factors (Figure 6a and 6b, and Supplementary Fig. 17).

A second wave of RA inducible gene activation was identified during the formation of neural progenitors. This is accompanied by transcriptional inactivation of *Stra8*, and activation of *Crabp1* and a second set of homeobox genes, *Hoxa2*, *Hoxa3*, *Hoxc5*, *Hoxd3* and *Hoxd4* (Figure 6a). This pattern of Hox gene activation suggests that the linear chromosomal arrangement of the Hox gene clusters does not necessitate temporal co-linearity in anterior Hox gene expression, a phenomenon that has been reported in the developing spinal cord[206]. Both waves of homeobox gene expression were accompanied by the upregulation of several long non-coding RNAs (lncRNAs) derived from the opposite strand (Figure 6b, Supplementary Fig. 17), consistent with previously identified lncRNA-based regulation of homeobox gene clusters[207-210].

The transition to mature neurons was marked by exit from the cell cycle and post-mitotic differentiation. In keeping with expectations, scTDA identified *Neurog1/2* (modulators of neuronal specification, cyclin regulation, cell cycle exit[211]) as well as *Ascl1* and *Sox9* among the cohort of known mediators of neuronal commitment, concomitant with a marked repression of topoisomerase 2A in the neuron population (Figure 6a).

We observed variability in eGFP expression levels and fluorescent intensities among differentiated neurons, suggestive of cellular heterogeneity at late stages of the differentiation. We performed a detailed analysis of this cellular population, and classified differentiated neurons into motor neurons (n = 343), and V1 (n = 19, *En1*$^+$), V2a (n = 10, *Vsx2*$^+$) and V2b interneurons (n = 15, *Gata3*$^+$) (Figure 6c), confirming the presence of cellular heterogeneity. Interestingly, among the differentially expressed genes between these populations we observed the presence of several lncRNAs. V1 interneurons uniquely express *Gm12688*, an intergenic lncRNA located near *Foxd3,* and transcribed from the opposite strand, which we validated using qPCR (Supplementary Fig. 18). Similarly, we identified a lncRNA located in chromosome 15 (NONCODEv4 accession number NONMMUG015572) as being uniquely expressed by V2b interneurons. Both V1 and V2b

GABAergic interneurons uniquely express *Gm14204*, an intergenic lncRNA located near *Slc32a1* and transcribed from the opposite strand. These results suggest a possible role of lncRNAs in neural diversification.

Among the genes associated with developmental transitions identified by scTDA, there were also multiple genes encoding RNA-binding proteins. These include known developmental state-dependent pre-mRNA splicing factors, as well as stage-specific but uncharacterized RNA binding proteins, which may guide cellular differentiation and post-mitotic commitment. In the context of the progenitor to neuron transition, our analysis identified *Nova1/2*, *Rbfox3*, *Srrm4*, and the Elav-like transcripts *Elavl4* and *Celf3* (Figure 6a), consistent with our expectations[212, 213]. scTDA further revealed the upregulation of *Mex3b* in the progenitor and post-mitotic cell populations, and constitutive expression of *Ptbp1* and *Ptbp2*. Previously published studies have documented *Srrm4* directed inclusion of neural specific exon 10 in *Ptbp1*[214]. Our results therefore suggest a transcriptional switch in splicing factor regulation that culminates in neural specific splicing.

scTDA as a General Approach to Interpret Heterogeneous Cellular Responses

The analysis and experimental validation presented above demonstrate that scTDA is an unbiased and scalable algorithm that chronologically orders asynchronous populations of single cells, while simultaneously preserving high-dimensional relationships between their transcriptional programs. We then asked if scTDA, applied to other data sets, could provide insights lacking in the previous analyses. To do so, we implemented scTDA to analyze three different *in vivo* cellular differentiations[215-217]. In each case, scTDA enabled an accurate reconstruction of developmental trajectories that extended the published analyses. First, we examined 80 single cells from the differentiating distal lung epithelium of mouse embryo[215]. All cells were sampled from the same time point (embryonic day E18.5). As shown in Figure 7a, scTDA recapitulated the proposed relationships between differentiating cells in both the bronchiolar and alveolar lineages. Moreover, our analysis uncovered a set of alveolar type I cells in which genes associated with mitochondrial respiration are down-regulated (Supplementary Fig. 19), suggestive of cellular stress or quiescence. This result shows that scTDA is also a valuable tool

for the analysis of one-time-point datasets from continuous asynchronous processes. Second, we applied scTDA to 1,529 cells from human pre-implantation embryos[216]. In this case, scTDA correctly ordered cells according to embryonic developmental time, and identified the segregation of several lineages, including the inner cell mass, the early trophectoderm, and a polar trophectoderm[216] (Figure 7b). In contrast to previous analysis, the identification and characterization of these cellular populations was completely unsupervised. Finally, we used scTDA to study 272 differentiating neurons in the mouse neo-cortex. Here, scTDA identified a continuum of cellular states with a bifurcation between apical and basal progenitors, and neurons (Figure 7c). The topological representation more accurately reflects a basal to apical progenitor migration and a split in potency, with apical progenitors sharing transcriptional profiles that more closely reflect their neuronal counterparts. This result highlights the ability of scTDA to faithfully represent nonlinear and converging cellular lineages, surpassing the capabilities of tree-based algorithms. From these results, we conclude that scTDA is an unbiased, generalizable and scalable approach to study differentiation at a single-cell level.

Discussion

We have developed a workflow that enables unsupervised analysis of high-dimensional single-cell developmental datasets. The biological context requires this approach to preserve the continuity of cellular differentiation, account for asynchronous development, and provide rigorous statistical interpretation of patterns of gene activation. We constructed a framework to accomplish this using topological data analysis (scTDA), and implemented it in a publicly available software. We applied this strategy to the *in vitro* differentiation of mESCs into neurons[2, 34]. This experimental and analytical approach strikingly recapitulated the continuous dynamic nature of cellular differentiation, and provided the necessary statistics to identify and characterize the transcriptional programs that accompany lineage restriction. Furthermore, scTDA revealed extensive transcriptional co-regulation of thousands of coding and noncoding genes expressed in precursor, progenitor and neuronal populations. Of particular note is the generality of scTDA, which can be applied to any biological system responding to inductive cues or environmental

perturbations (Figure 7). For example, scTDA may be used to study other cellular differentiation processes such as hematopoiesis, the evolution of cancer cells, neurodegeneration, and developmental disorders, all of which arise from extracellular signals and genomic alterations that culminate in heterogeneous transcriptional responses and cellular behavior.

Methods

Cell Culture and Single Cell Isolation

Murine embryonic stem cell (mESC) based differentiations were performed using the method of Wichterle *et al.*[34]. In brief, stem cell colonies are expanded on an adherent substrate, after which they are dissociated and monodispersed in a serum-free suspension (day 0). Individual stem cells aggregate into embryoid bodies (EBs), which maintain exclusive expression of pluripotency markers until they are induced down the neuron lineage by addition of RA and Smoothened Agonist (SAG) (day 2). Metabolized culture medium is replenished on day 5, coincident with the appearance of early eGFP positive cells within the EBs. EBs were dissociated into single cells using the Worthington Biochemical Papain Dissociation System (LK003178). Single cell deposition was accomplished using a Beckman Coulter MoFlo Astrios EQ cell sorter into 96 well plates. Cells were then snap-frozen and subsequently lysed.

Single Cell Library Generation

To obtain single cell expression profiles, modified CEL-Seq[124] was carried out. Briefly, we reverse transcribed the mRNA in each cell lysate with barcoded primers, pooled the single cell samples, synthesized second strand cDNA, and then linearly amplified by *in vitro* transcription with T7 RNA polymerase. The amplified RNA (aRNA) was chemically fragmented and T4 RNA ligase was used to ligate Illumina 3'-RNA adapters. The aRNA was then run on an Agilent Bioanalyzer to assess proper fragmentation and then reverse transcribed to generate cDNA and subjected to PCR enrichment. The subsequent multiplexed samples were paired end sequenced (2x125 bps) on an Illumina HiSeq 2500. Bulk RNA samples were purified from $2x10^6$ cells in 1 mL of Trizol

and standard Qiaquick RNA extraction protocols with a RIN 9.8 or higher. Stranded RNA-Seq libraries were generated at the New York Genome Center using a TruSeq Stranded Total RNA Library Prep Kit. Stranded cDNA libraries were paired end sequenced (2x125 bps) on a HiSeq 2500, operating in high output mode, yielding 30M reads per indexed library.

Immunofluorescence

EBs were washed three times in ice cold PBS and fixed for two hours at room temperature in 4% PFA. They were then washed with ice cold PBS and either embedded into OCT and stored at -80 C for sectioning, or stored in PBS at 4 C for whole EB staining. 20um sections were cut using a Leica CM 1950 cryotome and mounted onto a glass slide. Whole and sectioned EBs were incubated in blocking solution (0.1% tween-20, 5% donkey serum, 1% BSA) for two hours at room temperature followed by primary antibodies diluted in blocking solution overnight. Primary antibodies used were: goat-αSlc9a1 (Santa Cruz sc-16097, 1:100), rabbit-αOlig2 (Millipore AB9610, 1:100), mouse-αOlig2: clone 21F1.1 (Millipore MABN50, 1:100), rabbit-αEdnrb (Novus NLS54, 1:200), guinea pig-αVAChT (gift from Dr. Neil Schneider's lab, 1:400), rabbit-αSlc10a4 (Novus Biologicals NBP1-81134, 1:100), mouse-αPou5f1 (BD Biosciences 611202, 1:100), goat-αPECAM1 (Santa Cruz sc-1506, 1:100), sheep-αCD44 (R&D Systems AF6127, 1:250), goat-αFoxc1 (Santa Cruz sc-21396, 1:50), and goat-αSstr2 (Santa Cruz sc-11606 , 1:50). Alexa Fluor-conjugated secondary antibodies from Life Technologies were used at a 1:1000 dilution for two hours at room temperature. CD44, Foxc1 and Sstr2 were respectively conjugated to Cy-5, Alexa-405, and Cy-3 fluorophores using the DyLight system from Abcam (ab201798, ab188287, ab188288). Coverslips were mounted with Vectashield and EBs were imaged on an Olympus Fluoview FV1000 microscope using Olympus Fluoview v4.1.

Processing of Single Cell RNA-seq Data

Paired-end 125bp reads were de-multiplexed, trimmed, and mapped to the UCSC mouse reference (mm10) using Tophat[218]. Gene expression was quantified using transcript read counts as derived from HTSeq[219]. Read counts were normalized as,

$$e_i = \log_2\left(1 + \frac{10^6 \cdot r_i}{r}\right)$$

where r_i denotes the unambiguous read count for transcript i, and r denotes the total number of reads that are mapped to transcripts in the cell. A strategy based on spike-in read counts, as described by Stegle et al.[220], was implemented to filter out cells with a low content of mapped RNA and/or low sequencing depth. Specifically, the ratio ϵ_1 between average spike-in reads in the cell and average spike-in reads in the library (for spike-ins with an average of more than five reads in the library) was used to discard cells with low sequencing coverage. Similarly, the ratio ϵ_2 between the total number of spike-in reads in the cell and the total number of mapped reads was used to discard cells with a low number of mapped reads, relative to the sequencing coverage. Those cells showed very low expression of house-keeping genes, possibly representing cells under stressed conditions and/or with large amounts of degraded RNA. Based on the distribution of ϵ_1 and ϵ_2 in each of the two experiments (Supplementary Fig. 2a), cells with $\epsilon_1 > 5 \cdot 10^{-4}$ and $0.7 > \epsilon_2 > 0.01$ in the pilot experiment, and cells with $4.0 > \epsilon_1 > 0.05$ and $1.0 > \epsilon_2$ in the main experiment, were kept for subsequent analysis. The distribution of filtered out cells across different libraries was uniform in both experiments (Supplementary Fig. 2b). To reduce the noise observed near detection threshold in the pilot experiment, read counts with $r_i < 5$ were set to zero. Additionally, one of the libraries (RPI36) was discarded from subsequent analysis, because it presented a large batch effect. To assess the dependence of the library complexity on the differentiation time (Supplementary Fig. 3), we computed at each time point the distribution of the geometric library size, defined as the sum of log expression values over all genes in a cell[203].

Topological Representation

The algorithm Mapper[201] (Supplementary Note 1) was used to build topological representations of the RNA-seq data, through the implementation by Ayasdi Inc. Several open-source implementations are also available (https://github.com/MLWave/kepler-mapper, http://danifold.net/mapper/, https://github.com/RabadanLab/sakmapper, https://github.com/paultpearson/TDAmapper). In brief, the processed RNA-seq data was

endowed with a dissimilarity matrix by taking pairwise correlation distance (1 - Pearson correlation). To minimize the effect of drop-out events present in single-cell data, we only considered the 5,000 genes (for the pilot experiment) and the 4,600 genes (for the main experiment) with highest variance across each dataset. These are highly expressed transcripts for which the probability of not being captured by the RNA amplification (drop-out events) is small[221-224]. The space was reduced to $\mathbb{R}^2$ using MDS, as displayed in Figure 4c. A covering of $\mathbb{R}^2$ consisting of 26×26 (62×62) rectangular patches was considered for the pilot (main) experiment, respectively. The size of the patches was chosen such that the number of cells in each row or column of patches was the same, hence avoiding sampling density biases. The overlap between patches was 66% on average. Single linkage clustering was performed in each of the pre-images of the patches according to the algorithm described in Singh et al.[201]. A network was constructed where each vertex corresponds to a cluster, and edges correspond to non-vanishing intersections between clusters. We checked for the absence of batch effects in the topological representation (Supplementary Figs. 3d and 7) and the stability against different choices for the threshold for the number of genes used to compute the distance matrix and for the covering of $\mathbb{R}^2$ (Supplementary Fig. 12).

Gene Connectivity, Centroid and Dispersion within the Topological Representation

A notion of gene connectivity in the topological representation was introduced, defined as

$$s_i = \frac{N}{N-1} \sum_{\alpha,\beta \in \Gamma} \frac{e_{i,\alpha} A_{\alpha\beta} e_{i,\beta}}{\left(\sum_{\gamma \in \Gamma} e_{i,\gamma}\right)^2}$$

where $e_{i,\alpha}$ represents the average expression of gene i in node α of the topological representation, normalized as described in "Processing of RNA-seq" paragraph; Γ denotes the set of nodes of the topological representation; $A_{\alpha\beta}$ is its adjacency matrix; and N the total number of nodes. With this normalization, s_i takes values between 0 and 1.

The gene connectivity score depends on the distribution of expression values of the specific gene (Supplementary Fig. 10), and therefore genes cannot be ranked accordingly to their gene connectivity score in a meaningful way. To assess the magnitude of the connectivity score

relative to genes with the same expression profile, and rank genes accordingly, we introduced a non-parametric statistical test. We tested for the null hypothesis of a randomly expressed gene with the same distribution of expression values having a higher gene connectivity score. To that end, a null-distribution was built for each gene i using a permutation test. Cell labels were randomly permuted 5,000 times for each gene, computing s_i after each permutation. A *p*-value was estimated by counting the fraction of permutations that led to a larger value of s_i than the original one. Gene connectivity and its statistical significance were computed for each gene expressed in at least three cells. The resulting *p*-values were adjusted for multiple testing by using Benjamini-Hochberg procedure for controlling the false discovery rate.

To establish a pseudo-temporal ordering within the topological representation, a notion of root node was introduced. The latter was defined as the node that maximizes the function,

$$r(\alpha) = corr(d_\alpha, t)$$

where corr(*x*,*y*) denotes Pearson's correlation coefficient between *x* and *y*; d_α is the graph distance function to node α, that assigns to each node of the topological representation a value corresponding to the number of edges that are crossed in the shortest path from that node to node α; and t is the chronological sampling time function, that assigns to each node of the topological representation the average sampling time (expressed in days) of the cells contained in the node.

Least-squares linear regression was performed to determine the best fit for the coefficients a_0 and a_1 in the relation

$$d_{root} \cong a_1 t + a_0$$

where d_{root} is the graph distance function to the root node, determined in the previous paragraph. These coefficients were used to define the centroid and dispersion of each gene in the topological representation, expressed in days, and given respectively by

$$c_i = \frac{1}{a_1}\left(\frac{\sum_{\alpha\in\Gamma} d_{root}(\alpha) e_{i,\alpha}}{\sum_{\beta\in\Gamma} e_{i,\beta}} - a_0\right)$$

and

$$k_i = \frac{1}{a_1}\left(\sqrt{\frac{\sum_{\alpha\in\Gamma}(d_{root}(\alpha) - c_i a_1 + a_0)^2 e_{i,\alpha}}{\sum_{\beta\in\Gamma} e_{i,\beta}}} - a_0\right)$$

Such normalization, i.e. using coefficients a_0 and a_1 to express the centroid and dispersion in units of pseudo-time (days), allows comparing the connectivity and dispersion of a gene across different topological representations or studies.

Significance of Topological Features

We computed the first persistent homology group[225, 226] using the graph distance of the topological representation. Given the pairwise distances of a set of points sampled from a space, persistent homology allows to quantify the topological features (connected components, loops, cavities, etc., preserved under continuous deformations of the space) compatible with the data at a given scale. The first homology group, in particular, classifies loops of the space (Supplementary Note 1). We used persistent homology death times as a proxy of the size of the loops, and evaluated their statistical significance using a permutation test. To that end, we randomly permuted 500 times the labels of the genes, for each cell independently. For each permutation we built a topological representation using the same parameters than in the original representation and computed the first persistent homology group. A *p*-value for each of the loops was estimated from the distribution of the number of loops as a function of their death time. The resulting *p*-values were adjusted for multiple testing by using Benjamini-Hochberg procedure for controlling the false discovery rate.

Comparison to Other Methods for Analyzing Longitudinal Single-Cell RNA-seq Data

We dimensionally reduced the processed single cell RNA-seq data of the main experiment using MDS, ICA and tSNE, using the same set of highly expressed, highly variant genes that we used for building the topological representation. In each representation we determined the cell that maximized the Pearson correlation coefficient between the two-dimensional Euclidean distance to the cell and chronological sampling time, corresponding to the least differentiated cellular state.

Additionally, we compared scTDA to the single cell software Monocle[181], based on ICA and minimum-spanning trees, Wishbone[185], based on diffusion coefficients, and SLICER[184], based on locally linear embedding. We followed all recommendations in the documentation of these algorithms. In our tests, Monocle failed in running over the complete main experiment dataset, consisting of 2,304 cells, and only a partial set of 834 cells, sampled from all time points was analyzed.

Simulated Data

Noisy, branched asynchronous cellular differentiation processes were simulated, from which 700 cell were sampled at three time points. To that end, we used the following strategy:

1.- We simulated a noisy branched tree-like structure with three parameters (*t*, *u*, and *v*), and performed a non-linear transformation into four-dimensional embedding space (spanned by the variables $g^{(1)}$, $g^{(2)}$, $g^{(3)}$, and $g^{(4)}$). This space provides the structure for four different groups of correlated genes.

2.- We randomly sampled 700 points from this space, corresponding to 700 cells, and assigned a sampling day based on a multinomial distribution with probabilities given by a logistic function of *t*, to simulate asynchrony.

3.- We simulated the expression of 300 genes driven by the variables $g^{(1)}$, $g^{(2)}$, $g^{(3)}$, and $g^{(4)}$, in addition to 200 genes with non-correlated expression, sampled from normal distributions.

4.- We simulated the effect of drop-out events using the standard logistic dependence of the drop-out probability as a function of expression.

In what follows, we provide a detailed description of each of these steps.

First, we simulated four groups of correlated genes. These were defined by the equations,

$$g_i^{(1)} = \frac{200\, u_i}{u_i^2 + t_i^2 + 0.2} + N_i^{(1)}$$

$$g_i^{(2)} = \frac{200\, t_i}{u_i^2 + t_i^2 + 0.6} + N_i^{(2)}$$

$$g_i^{(3)} = \frac{100(u_i^2 + t_i^2) - 20}{u_i^2 + t_i^2 + 0.2} + N_i^{(3)}$$

$$g_i^{(4)} = 100\sqrt{v_i} + N_i^{(4)}$$

where $N_i^{(k)}$, are normally-distributed random variables with mean $\mu = 150$ and standard deviation $\sigma = 8$. The index $i = 1, \dots, 700$ runs across the sampled cells, and u_i, v_i, and t_i are randomly sampled from

$u_i = 0, \quad v_i = 0 \qquad 0 \le t_i < 0.2,$ or

$u_i = 0.2 - t_i, \quad v_i = 0 \qquad 0.2 \le t_i < 0.7,$ or

$u_i = t_i - 0.2, \quad v_i = 0 \qquad 0.2 \le t_i < 0.7,$ or

$u_i = 0.6 - t_i, \quad v_i = 0 \qquad 0.4 \le t_i < 0.7,$ or

$$u_i = -0.2, \quad v_i = t_i - 0.4 \qquad 0.4 \le t_i < 0.7,$$

where only the first $2 + r$ equations are considered in simulated differentiation processes with r lineage branching points. The variable t_i represents the differentiation pseudo-time of cell i at the time of sampling. To simulate asynchrony, each sampled cell was assigned a sampling day accordingly to

$1 \ge p_i > \tau(t_i, 0.23)$ → day 1

$\tau(t_i, 0.23) \ge p_i > \tau(t_i, 0.47)$ → day 2

$\tau(t_i, 0.47) \ge p_i \ge 0$ → day 3

where $p_i \in [0,1]$ is a random number uniformly distributed, and τ is the logistic function

$$\tau(a, b) = \frac{1}{1 + e^{15(b-a)}}$$

We simulated 75 genes in each of the four groups of genes, with expression values given by

$$m_{l,i}^{(k)} = g_i^{(k)} N'^{(k)}_l N''^{(k)}_i, \qquad k = 1, \dots, 4, \qquad l = 1, \dots, 75, \qquad i = 1, \dots, 700$$

where $N'^{(k)}_l$ and $N''^{(k)}_i$ are normally-distributed random variables with mean $\mu = 1$ and standard deviation $\sigma = 0.2$. In addition, we simulated 200 extra genes with non-correlated expression

$$m_{l,i}^{(0)} = N'^{(0)}_l N''^{(0)}_i, \qquad l = 1, \dots, 200, \qquad i = 1, \dots, 700$$

with $N'^{(0)}_l$ and $N''^{(0)}_i$ normally-distributed random variables with mean $\mu = 200$ and 1, and standard deviation $\sigma = 50$ and 0.2, respectively. Hence, a total of 500 genes were simulated in

each of the 700 cells. To model the effect of drop-out events, we randomly set to zero the expression of some of the genes in some of the cells, with probability

$$P = \frac{1}{1 + e^{m-1}}$$

where m is the original expression value of the gene in the cell.

Gene Ontology Annotation

Gene ontologies were obtained from EMBL-EBI QuickGO[227]. Specifically, categories GO:0006355 "Regulation of transcription, DNA-templated", GO:0008380 "RNA splicing", GO:0044822 "Poly(A) RNA binding", GO:0051726 "Regulation of cell cycle", and GO:0007049 "Cell cycle" were used to annotate genes. Expression of genes associated to DNA replication was analyzed by considering the 99 genes in the category GO:0006260 "DNA replication" expressed in less than 1,400 cells in the main experiment. Genes coding for proteins in the cellular surface were identified by looking in UniProt database[228] for proteins annotated with an extracellular topological domain. Gene ontology enrichment analysis was performed using PANTHER classification system[229].

Transient Cellular Populations

Low-dispersion genes ($k_i < 1.7$ days and $k_i < 2.25$ days respectively in the main and pilot experiments) with significant gene connectivity ($q < 0.05$) in the topological representation where clustered according to their centroid using k-means clustering (Supplementary Fig. 13b and 13c). The optimal number of clusters according to Davies-Bouldin index was four in the main experiment (three in the pilot experiment) (Supplementary Fig. 13a), as it was also evidenced from visual inspection of the centroid distribution for low-dispersion genes. A state $r = 1,\ldots,4$ was assigned to each node of the topological representation based on the average expression of each cluster of low-dispersion genes in the node (Supplementary Fig. 13d). Genes with significant gene connectivity according to the permutation test described in the paragraph "Gene Connectivity, Centroid, and Dispersion within the Topological Representation" were assigned to

each of the four populations based on the number of cells expressing the gene in each state *r*. Only genes expressed in at least 80 cells and at most 1,500 cells were considered.

Analysis of Long Non-Coding RNAs

The coordinates of intergenic and antisense lncRNAs were downloaded from NONCODEv4[230], and read counts were obtained using HTSeq. The connectivity and statistical significance of each long non-coding gene in the topological representation was computed using scTDA. Only lncRNAs that were significant ($q < 0.05$) in both the pilot and main experiment, and that were supported by at least 50 reads in the longitudinal stranded RNA-seq data were kept. Curation was performed to remove lncRNAs whose 3'-end overlapped the 3'-end of another gene, read assignment therefore being ambiguous, or that corresponded to possible miss-annotations of the 3' UTR of a nearby gene.

Characterization of Interneuron Populations

Differential expression analysis between *En1*$^+$, *Gata3*$^+$, *Vsx2*$^+$ and *Egfp*$^+$ cells in nodes characterized as post-mitotic (Supplementary Fig. 13d, state 3) was performed using the software SCDE[41] with default parameters.

scTDA Software

The algorithms described in this work were implemented and documented in an object oriented python library for topological data analysis of high-throughput longitudinal single-cell RNA-seq data, called scTDA. scTDA is publicly available at https://github.com/RabadanLab/scTDA.

Online Database

We developed a database that allows easily exploring the topological representations and statistics of the two motor neuron differentiation datasets. The database is publicly available at https://rabadan.c2b2.columbia.edu/motor_neurons_tda.

CHAPTER 3: Modeling Amyotrophic Lateral Sclerosis *in vitro*

Amyotrophic Lateral Sclerosis is a paralytic disease that selectively impacts subpopulations of motor neurons. During the onset and progression of ALS, each cell type within the spinal cord responds and contributes to a toxic effect on motor neurons[10]. Whole exome and genome sequencing has identified familial and sporadic mutations in genes associated with ALS, and these genes are transcriptionally active within both neurons and glia. Mechanistic insights into disease progression have been provided by *in vivo* and *in vitro* modeling with such ALS-associated genes. The most accurate ALS mouse model is based on overexpression of the superoxide dismutase 1 (SOD1) gene bearing a mutation shown to cause familial ALS. Introducing an ALS-associated variant of SOD1, such as $SOD1^{G93A}$, into model organisms phenocopies ALS disease progression[11]. *In vivo* mouse models have revealed the contributions of glia and neurons to the onset, and progression of ALS by the conditional expression of $SOD1^{G93A}$ in specific cell types[12-16]. Astrocytes, which normally function as support cells for neurons, become reactive and secrete toxic moieties, inducing and accelerating motor neuron degeneration[17-20]. Microglia, the macrophage-like immune cells of the central nervous system, become activated and release pro-inflammatory signaling molecules[21-23]. Mature oligodendrocytes are replaced by an immature population lacking the capacity to myelinate axons[24, 25]. Taken together, these observations motivate a deeper understanding of cell type specific effects on ALS disease progression. These complex pathological events can be modeled more simply with *in vitro* cultures of defined cells, which can be generated from primary cells isolated from animal models or from human patients. Primary astrocytes, microglia, and cortical neurons can be expanded in culture and studied directly, while motor neurons are commonly obtained through reprogramming stem cells. Although limited in their in vivo predictive potential, *in vitro* models offer a well-defined, scalable approach to monitor and test hypotheses of ALS disease progression. Ensemble averaged transcriptomic studies performed *in vitro* using sandwich cultures have identified key pathways triggered by cell autonomous and cell non-autonomous effects[65]. We sought to disentangle the extrinsic effect of glia from the intrinsic

alterations occurring within individual motor neurons during ALS disease progression, seeking to resolve dynamic events and subpopulations masked in ensemble averaged measurements.

Chapter two of this thesis provided a computational and experimental paradigm to capture the continuous structure associated with single cell studies of heterogeneous cellular responses. In that work, topological data analysis coupled with single cell sequencing enabled an unsupervised dissection of the transcriptional programs and accompanying regulatory architecture associated with the *in vitro* differentiation of mouse embryonic stem cells into motor neurons. Given an abundance of *in vivo* knowledge available, in vitro standardization, and stereotypy in results, this system provided several advantages for a proof of concept study. Furthermore, the motor neurons generated from chapter two are often utilized to model ALS in vitro, supporting their use in feasibility studies. Although that workflow provided a meaningful approach to the study of continuously structured cellular responses, further computational and experimental innovation was required to its application towards the study of ALS disease progression.

As mentioned in the introduction, subpopulations of a, b, and g motor neurons may be defined by transcriptional signatures[6]. Despite interest in the identification of new markers for motor neuron identity, single cell sequencing is limited by its depth of coverage, limiting further refinement of motor neuron subclasses. In addition, ensembles of highly similar groups of individual cells are difficult to classify using traditional similarity metrics such as linear correlation. In this study, information theory has been linked with topological data analysis to provide a detailed insight into the dynamics associated with, and the relative contribution of, cell autonomous and non-autonomous effects on gene expression in motor neurons during ALS disease progression. Although the results from this study are complex, I will focus the discussion on three emergent principles identified; glial contributions to accelerated neuronal aging, altered nutrient homeostatic signatures, and counteraction against innate immunity.

Among the methods employed to classify cell types from single cell sequencing, Mutual information has been used previously to analyze and interpret subpopulations of projection neurons in Drosophila Melanogaster. In that study, genetic identity was tied to patterns of

projection to the mushroom body of the fly brain[231]. Further computational studies demonstrated that mutual information, defined as the mutual dependence of two random variables on one another, is capable of capturing nonlinear and non-monotonic dependencies between genes or groups of genes[232]. A characteristic metric associated with mutual information is known as the Jaccard index, which encapsulates entropic similarity between genes. Rather than utilize Pearson correlation as a metric for similarity across the transcriptomes of individual cells (as performed in scTDA), we utilized the Jaccard Index to generate topological representations from motor neurons undergoing disease progression.

Stem cell based models of neurodegenerative disease provide a platform to study disease progression in a well-defined environment. As a corollary, unique contributions of individual genes and cell types to disease may be elucidated. Unlike primary motor neurons, which are difficult to isolate and culture, stem cell-derived motor neurons allow hundreds of thousands of cells to be profiled at once. *In vitro* models are easily manipulated, and can be purposed to studying genetic variants introduced via gene editing, silencing, or overexpression. Furthermore, cells can be co-cultured in different combinations, allowing cell type-specific contributions to ALS to be identified. Murine stem cell (mESC) derived motor neurons (ESMN), human stem cell (hES), and human induced pluripotent stem cell (hiPSC) derived motor neurons (hES-MN and hiMN) have been co-cultured with astrocytes, microglia, and myocytes to study ALS disease progression[233]. hES-MN and hiMN models have the advantage of being human-based and patient-specific, thereby allowing modeling of sporadic ALS and the full spectrum of genetic variation in the background of the genome of an ALS patient, However, these models do not replicate the ultimate degenerative phenotype of ALS in vivo. Mice that overexpress the human ALS-causing $SOD1^{G93A}$ mutation have been used to model ALS since 1994[11], and ESMNs differentiated from their mESCs have been used to study ALS *in vitro*.

ESMNs constitute a FACS-purified population of Hb9::eGFP positive cells from dissociated EBs at day 7 of differentiation. They can be generated from mESCs harboring any mutation, allowing for interrogation of any gene implicated in ALS for which a mouse model is available. Furthermore, non-transgenic mESCs can be genetically manipulated to express a

reporter gene or genetic mutation, providing an alternative to mouse models. ESMNs display transcriptional programs consistent with LCMD motor neurons from the spinal cord[65], and morphologically appear like primary spinal motor neurons. Based on the expression of Hox genes, which determine rostro-caudal positioning of motor neurons in the spinal cord, the majority of ESMNs most highly resemble cervical motor neurons[34]. *in vitro* ESMNs require neurotrophic factors from astrocytes to survive in culture, and are plated either in sandwich culture or co-cultured with primary astrocytes isolated from neonatal mice. In control models, these astrocytes come from nontransgenic (NT) or WT hSOD1 (WT) overexpressing mice, while in disease models astrocytes are derived from ALS model animals. ESMNs overexpressing $SOD1^{G93A}$ exhibit decreased viability compared to $SOD1^{WT}$ ESMNs and accelerated degeneration when cultured in the presence of $SOD1^{G93A}$ glia. Astrocyte toxicity has been shown to be ESMN specific, as mESC-derived interneurons and ocular neurons display consistent survival in the presence or absence of $SOD1^{G93A}$ expressing cells.[51, 234] Recently human reprogrammed astrocytes have been used in coculture with ESMNs to study the effects of patient astrocytes on motor neuron health and viability[235].

A timecourse study over a two week $ESMN^{G93A}$ – murine $SOD1^{G93A}$ astrocyte sandwich culture has revealed transcriptional changes in astrocytes and ESMNs that are ALS specific.[65] These effects can be cell autonomous, taking place in $SOD1^{G93A}$ ESMNs in the presence of WT astrocytes, or non-cell autonomous changes that are induced in ESMNs by the presence of a secreted factor from mutant astrocytes. Among these are changes in TGFb signaling, which have been validated in in vivo, in the murine spinal cord. Axonal degeneration is one of the first signs of ALS[236]. *in vitro* Axon-Seq has shown differences in the transcriptional profiles of somas of $SOD1^{G93A}$ ESMNs and their axons. Over a hundred of these genes are dysregulated in the axons of $SOD1^{G93A}$ ESMNs, corresponding to both pathological and compensatory programs[237].

Although motor neurons in the spinal cord are selectively vulnerable to degeneration in ALS, not all spinal motor neurons are equally affected. Fast motor neurons, marked by the expression of Matrix Metalloprotease 9, are highly vulnerable.[238] Comparative studies of vulnerable neurons with resistant ones has further implicated genes involved in protein transport

and excitability in marking motor neurons as differentially vulnerable or resistant.[239] Given the differences in motor neuron disease susceptibility *in vivo*, and the differential *in vitro* viability of ESMNs, it is possible that there exists an unappreciated underlying heterogeneity in ALS disease progression in cultured ESMNs. This heterogeneity is obscured in bulk data, but can be resolved with scRNA-seq. A timecourse study with increased resolution of $SOD1^{WT}$ and $SOD1^{G93A}$ ESMNs co-cultured with $SOD1^{WT}$ and $SOD1^{G93A}$ murine astrocytes will allow the interrogation of cell specific pathway activation, determine the fractional composition of cells expressing a given transcript, and identify outlier populations that may contaminate bulk data with strong expression.

Results

ESMNs differentiated from $SOD1^{WT}$ and $SOD1^{G93A}$ mESCs were cultured over $SOD1^{WT}$ and $SOD1^{G93A}$ primary murine astrocytes. At days 3, 7, and 14 post-plating, cells were dissociated and eGFP+ ESMNs were FACS-sorted into 384 well plates. scRNA-seq libraries were generated using a modified SCRB-seq protocol, with 8500 genes detected per cell on average (Supplementary Figure 20). The following annotation will be used do identify ESMN genotype and culture condition:

WT/WT = $SOD1^{WT}$ ESMNs cultured over $SOD1^{WT}$ astrocytes
WT/G93A = $SOD1^{WT}$ ESMNs cultured over $SOD1^{G93A}$ astrocytes
G93A/WT = $SOD1^{G93A}$ ESMNs cultured over $SOD1^{WT}$ astrocytes
G93A/G93A = $SOD1^{G93A}$ ESMNs cultured over $SOD1^{G93A}$ astrocytes

Visualization of transcriptional profiles of single ESMNs using kNN

As a group, ESMNs have a highly similar transcriptional program regardless of the conditions they are sampled. This is because ESMNs have been FACS purified to ensure they are Hb9::eGFP positive post-mitotic cells. The most direct approach to find variability in a highly homogeneous population is through PCA followed by dimensional reduction using Euclidean distance. Cells (nodes) are plotted on a k-nearest neighbor (kNN) graph using a force-directed

layout[240]. Edges in the graph represent connections between neighboring cells in high-dimensional space, however most of these edges have been removed during visualization and can be excluded from the display without compromising data integrity (Figure 8). As with tSNE plots, the physical absolute coordinates of the cells have no intrinsic meaning beyond visualization, while orientation of cells relative to one another is biologically meaningful. Transcriptional variability will be observed not just between cells sampled under different conditions, but between cells sampled at different time points. Because the algorithm is not focused on maintaining continuity of the data, but rather identifying differences in transcriptional profiles, the timecourse is not linear. However, the WT/WT condition at days 3, 7, and 14 can be taken as a baseline for normal aging of ESMNs in culture. The clusters identified by kNN do not show complete exclusivity for conditions and sampling time. For example, instead of forming a cluster of their own, WT/G93A cells at day 7 are divided between the day 7 WT/WT and G93A/WT clusters.

Clustering of ESMNs with kNN

For each cluster, the predominant ESMN populations are outlined below. When fewer than 10 cells from a different sample are present, they are excluded from the labeling (Table 1). Force directed layouts afford flexibility such that clusters can be relocated to a position more consistent with the remainder of the graph without altering the information content. The clusters are defined in three principal ways: 1) the expression level of a transcript, 2) the number of cells that express a transcript, 3) uniqueness of a transcript to a given cluster.

ESMNs sampled at day 3 form 4 distinct but largely continuous clusters in the kNN graph (clusters 1-4). The transcripts responsible for segregating the samples are nonexclusive to the cluster, but rather are expressed at higher levels in subsets of ESMNs. Cluster 3 appears more separated from the others. This is due to a higher sampling depth, not to expression of unique transcripts.

Cluster 1: Day 3 WT_WT

Cluster 2: Day 3 WT_G93A

Cluster 3: Day 3 G93A_WT

Cluster 4: Day 3 G93A_G93A

ESMNs sampled at day 7 form 3 distinct but continuous clusters in the kNN graph (clusters 5-7). Again, the clustering is weighed by expression level more than unique transcripts. The expression level differences that started in day 3 become more pronounced at day 7 (eg, Gm1821, Selm). New transcripts also appear as differential between clusters (eg, S100a11, Gpx3, Ddx42).

Cluster 5: Day 7 WT_WT; D7 WT_G93A; D3 G93A_G93A

Cluster 6: Day 7 G93A_WT; Day 7 WT_G93A

Cluster 7: Day 7 G93A_G93A; Day 7 G93A_WT

ESMNs sampled at day 14 form 4 main clusters and 2 less populated clusters which are subsets of two of the main ones. Again, within clusters there are cells sampled at different conditions. There are now concrete differences in the transcripts that are expressed between clusters (Hspb2, Saa3, Mlkl), the number of cells in each cluster that express a given transcript (Pvalb), and the expression level of the transcript (Ftl1).

Cluster 8: Day 7 G93A_G93A, Day 14 WT_WT

Cluster 9: Day 14 WT_WT, Day 14 WT_G93A

Cluster 10: Day 14 WT_G93A; Day 14 G93A_WT

Cluster 11: Day 14 G93A_WT; Day 14 G93A_G93A

Cluster 12: Day 14 WT_WT, Day 14 WT_G93A

Cluster 13: Day 14 G93A_WT; Day 14 G93A_G93A

Dying populations

At day 14, two populations of ESMNs strongly enriched for stress markers (Lyz1, C1qa, C1qb, and C1qc) appear (clusters 12 and 13, Supplemental Figure 22A). The first (cluster 12), projecting from a predominantly day 14 WT/G93A cluster, is composed of 9 WT/WT and 28

WT/G93A ESMNs and corresponds to non-cell autonomous cell death pathway. The second (cluster 13), projecting from cluster 11, is composed of 11 G93A/WT and 48 G93A/G93A ESMNs, corresponding to a cell autonomous cell death population. Expression of genes spanning several pathways, including cytoskeletal genes such as Cnn1 and Acta2, segregate the two populations. Between clusters 11 and 13, the expression level of stress markers in cells is a continuum such that the cells in cluster 11 most distant from cluster 13 have the least expression of stress markers while the cells more adjacent to cluster 13 have a stronger expression of stress markers. The same pattern of expression can be seen between clusters 10 and 12, however the dropoff of stress markers is more pronounced in the distal cells of cluster 10.

Shared profiles of ESMNs between sampling days

Excluding subgranular and subventricular zones, within the mammalian body neuronal cells are resident for the lifespan of the animal. Interestingly, the data indicate a surprising accelerated aging process in $SOD1^{G93A}$ neurons co-cultured with mutant glia. While $SOD1^{G93A}$ results in distinct transcriptional changes in ESMNs both cell autonomously and non-cell autonomously, it also results in a previously unappreciated acceleration of aging in a subset of G93A/G93A ESMNs. Out of 383 G93A/G93A ESMNs sampled at day 3, 26 (7%) clustered with WT/WT ESMNs sampled at day 7. This trend becomes more apparent at day 7, when 51 out of 363 (14%) G93A/G93A ESMNs exhibit a transcriptional program that results in them clustering with WT/WT ESMNs sampled at day 14.

Continuum of transcriptional changes associated with ALS

Overlap between ESMNs sampled at different conditions shows a continuity in ALS-associated transcriptional changes. At timepoints 7 and 14, each cluster is composed of a mix of cells representing a combination of either: 1) WT_WT and WT_G93A; 2) WT_G93A and G93A_WT; 3) G93A_WT and G93A_G93A. Interestingly, these combinations are consistent – WT_WT ESMNs do not appear in the same cluster as G93A_WT or G93A_G93A ESMNs. This implies a hierarchy governing cell autonomous and non-cell autonomous effects. The non-cell autonomous

contributions of SOD1^{G93A} astrocytes can induce a transcriptional program in a subset of SOD1^{WT} ESMNs that change their transcriptional profiles to more closely resemble those of SOD1^{G93A} ESMNs cultured with SOD1^{WT} glia. However, this effect is not sufficient to reproduce the "disease" transcriptional program of G93A_G93A ESMNs. This is time-dependent, as day 3 ESMNs form distinct clusters that are delineated by the condition under which they were sampled.

Expression of non-neuronal genes in ESMNs sampled at day 14

Surviving ESMNs at day 14 start displaying transcriptional signatures of astrocytes, microglia, and oligodendrocytes (Figure 9). We limit our analytical focus to earlier timepoints, given the preponderance of motor neuron death at later stages and the subsequent enrichment of non-neuronal cell type contaminants. That said, such non-neuronal transcriptional signatures have previously been shown to occur both *in vivo* and *in vitro* studies of neurons in disease and health.[241-244] It is possible that the expression of these RNAs in ESMNs are a consequence of artificial culturing conditions, as they have previously been seen in transcriptional studies of ESMNs[65]. Likewise, it is possible that the presence of SOD1^{G93A} promotes the expression of these genes in late-stage ESMNs. It has also been shown that astrocytes secrete extracellular vesicles (EVs) containing miRNAs that have a functional on ESMNs.[245-247] It is possible that the altered transcriptional programs of late-stage ESMNs is impacted by exposure to EVs containing regulatory elements from support cells.

Single Cell Topological Data Analysis applied to post-plated ESMNs

Clustering approaches such as kNN can provide strong insights into the transcriptional differences between groups of cells, but lose resolution for continuity in the data. scTDA coupled with mutual information allows the determination of a "map" of transcriptionally continuous data in a background of high similarity. In order to gain insights into both cell autonomous contributions of SOD1^{G93A} in ESMNs and the non-cell autonomous contributions of SOD1^{G93A} primary astrocytes, as well as cumulative changes occurring in ALS disease, I generated three separate topological

representation of the timecourse data using only the cells sampled at days 3 and 7, which show high expression of pan-neuronal markers (Figure 10, Supplemental Figure 23). For generating the representations, I used ~1500 genes of maximum variance. One key difference between these representations and the one in Chapter 2 is the circular structure found in autonomous and disease states (Figure 10A, C). This is a result of continuity in the data from shared transcriptional programs between cells sampled at different conditions. In combination, ESMN genotype and culture time result in loops, or holes, in the representation, while nonautonomous astrocyte contributions result in a spreading of time-ordered data. Biologically, this structure suggests that the ESMN genotype and *in vitro* culture time have a greater impact on the transcriptional readout of ESMNs than the extrinsic effects of $SOD1^{G93A}$ glia. It is important to remember that in scTDA representations, greater physical distances (longer edges) in the graph do not correlate with greater transcriptional differences, but are rather a function of visualization. Distance between nodes is calculated as the fewest number of edges connecting two nodes.

The first is a comparison of cell autonomous changes: $SOD1^{WT}$ and $SOD1^{G93A}$ ESMNs cultured over $SOD1^{WT}$ astrocytes (Figure 10A). The structure shows of a time-dependent bifurcation the data. Starting from closely related day 3 $SOD1^{WT}$ and $SOD1^{G93A}$ ESMNs, which occupy adjacent and overlapping nodes, the representation branches into two connected day 7 populations demarcated by ESMN genotype. The nodes connecting two populations of ESMNs at day 7 imply shared transcriptional programs in subpopulation of cells within those nodes, while the nodes connecting the day 7 populations to their respective day 3 populations retain transcriptional similarity to the earlier timepoints. The cells within these nodes can be considered transitioning cells, and we are in the process of examining the transcriptional programs that they encompass to understand the basis for asynchronous disease progression.

The second representation is a comparison of non-cell autonomous changes in $SOD1^{WT}$ ESMNs cultured over $SOD1^{WT}$ and $SOD1^{G93A}$ astrocytes (Figure 10B). In this representation, the cells are still time-ordered, however both the day 3 and the day 7 populations are adjacent and overlapping without loops in the structure. Interestingly, the ESMNs cultured over $SOD1^{G93A}$ astrocytes have very few direct connections between day 3 and day 7 samples, instead

transitioning through the late day 3 and early day 7 ESMNs cultured over $SOD1^{WT}$ astrocytes. This is consistent with accelerated aging that, in the kNN clustering, was called out in the disease condition.

The third representation is a comparison of disease changes in $SOD1^{WT}$ and $SOD1^{G93A}$ ESMNs cultured over $SOD1^{WT}$ and $SOD1^{G93A}$ astrocytes (Figure 10C). This representation shows not bifurcation in the data, but rather condensation of $SOD1^{WT}$ and $SOD1^{G93A}$ conditions over time. The starting populations at day 3 are largely distinct, with connecting nodes containing subsets of cells that are transcriptionally similar. Over time, the transcriptional programs coalesce into a largely connected day 7 population. One reason for this may be that that the surviving $SOD1^{G93A}$ neurons at day 7 in the disease model show greater transcriptional similarity to the $SOD1^{WT}$ neurons, while the selectively vulnerable population has degenerated. To understand the connections between the disease and the control conditions, we are again in the process of examining the transcriptional programs activated in nodes that connect populations.

Accelerated aging in $SOD1^{WT}$ ESMNs under stress of $SOD1^{G93A}$ astrocytes

Astrocytes provide neurons metabolic substrates and precursors utilized for energy production, neurotransmitter synthesis, along with lactate and glutathione regulation. Furthermore, astrocytes remediate neuronal accumulation of oxidized products and glutamate[248]. We sought to understand how diseased astrocytes contribute to oxidative stress within motor neurons and the resulting impact on canonical aging signatures. Thioredoxin interacting protein, Txnip, is upregulated across multiple species during aging, and is one of 9 genes that is transcriptionally upreguated in ten regions of the human brain during aging. Furthermore, it has been shown to be responsible for accelerating aging in cells and organisms[249, 250]. It functions through suppressing the actions of thioredoxin (Txr, encoded by the cytoplasm and nuclear localized Txn1 and mitochondrial localized Txn2), an antioxidant protein that promotes longevity through protective mechanisms against oxidative stress[251]. While both Txnip and Txr are constitutively expressed, overexpression of Txnip is induced in neurons by oxidative stress and contributes to neuronal apoptosis[252]. In these data, we found Txnip upregulation in subsets of

Day 3 $SOD1^{WT}$ and $SOD1^{G93A}$ ESMNs cultured over $SOD1^{G93A}$ astrocytes, but not in $SOD1^{G93A}$ ESMNs cultured over $SOD1^{WT}$ astrocytes, implying a non-cell autonomous role for the early activation of this gene (Figure 11). In congruence with this observation, Txr levels are low at the early timepoint and rise at day 7 in the nonautonomous and disease conditions, while remaining consistent in the cell autonomous condition, further implicating the contribution of astrocytes to the control of the Txr/Txnip pathway. We cross validated this phenomenon with spatial transcriptomics data collected from $SOD1^{G93A}$ spinal cords from mice[253]. The results are in agreement, with elevated P30 $SOD1^{G93A}$ observed within the spinal cord when compared to P30 $SOD1^{WT}$ (Supplementary Figure 26A). Oxidative stress is implicated as a key mechanism for motor neuron death in ALS, and understanding the early responses of motor neurons to oxidative damage, both through intrinsic and extrinsic signaling pathways, is important for further understanding disease onset and progression[254]. Outside of the central nervous system, Txnip dissociation from Txr accelerates inflammasome activation through its subsequent binding with the Nod like receptor protein 3, Nlrp3, and further provides a mechanistic switch between an adaptive versus terminal unfolded protein response[255]. The phenomenon, points to a potentially novel early stage mechanism that has gone unappreciated in motor neuron disease.

Dysregulation of iron homeostatic genes

Nutrient homeostasis is classically defined by a cell's ability to dynamically maintain a constant intracellular concentration of a given nutrient. Clearly, neurons are relatively unique in that they must cope with dramatic intracellular changes in calcium, sodium, potassium, and glutamate concentrations. These alterations, of course, are integral to the electrical activity occurring within neurons, which mediates synaptic communication. Supporting this activity are cell autonomous buffering mechanisms, such as calcium buffering by the endoplasmic reticulum[256]. Furthermore, astrocytes mediate glutamate clearance and provide nutritive support to neurons within the mammalian spinal cord[257]. Given our interest in cell autonomous and non-cell autonomous changes contributing to disease progression, we sought to identify alterations in nutrient homeostatic mechanisims during our time course experiments. We hypothesized that changes in

nutrient availability, namely availability and usage, may lend deep insights into changes in neuronal survival and susceptibility. Iron homeostasis is of critical importance to neuronal function, as iron is a critical cofactor in many enzymes responsible for electron transfer in redox reactions, including proteins involved in the electron transport chain for respiration, control of gene expression, and DNA repair.[258] Excess iron is also a source of free radicals and oxidative damage in cells.[259] Iron homeostasis is dysregulated in a number of neurodegenerative conditions, resulting in the accumulation of iron deposits in neurons.[260]

Biologically, iron is present in two forms: ferric iron (Fe^{3+}) and ferrous iron (Fe^{2+}). Iron is transported and stored in the ferric state, however metabolically it is used in the ferrous state. Iron is imported into cells through two pathways, one for ferric iron involving the transferrin receptor (Tfrc), the other for ferrous iron through a divalent metal ion transporter (eg, Slc11a2 and Slc39a14). While transferrin is not expressed in the sequencing data as it is not produced by neurons, it is present to the culturing medium through N2 supplement (see Methods). The transferrin receptor, Tfrc, is highly expressed in days 3 and 7 of culture, however expression is tapered in day 14 ESMNs, especially those in the diseased state corresponding to clusters 11 and 13. This is consistent with the expression of Tfrc in the spinal cord of $SOD1^{G93A}$ mice, as seen by spatial transcriptomics through ALS-ST.NYGENOME.ORG (Supplemental Figure 26B) [253]. Expression of Slc11a2 and Slc39a14 is ubiquitous throughout the timecourse, but Slc11a2 is more sparsely represented in clusters 3, 5 and 10 (Supplemental Figure 22B). Decreased expression of iron import proteins indicates diminished uptake of iron into ESMNs and surplus iron in these cells.

In order to better understand the trends in expression of the iron transport genes, we looked at the expression of the ferric iron transporter Tfrc and the ferrous iron transporter Slc11a2 in the scTDA representations (Figure 12). Tfrc and Slc11a2 are constitutively expressed across all samples, but their expression levels vary between conditions. Tfrc has higher expression at day 7 than day 3 in all conditions except for the day 3 nonautonomous ESMNs cultured over $SOD1^{G93A}$ astrocytes, which has comparable expression to day 7 timepoints. This is consistent with the "accelerated aging" of these ESMNs observed in the previous section. Slc11a2

expression is consistent in all samples of the non-cell autonomous condition. In day 7 of the $SOD1^{G93A}$ ESMNs in the disease condition, Slc11a2 is down-regulated while the expression of Tfrc is upregulated, suggesting a switch in the iron transport system. Tfrc has also been shown to be upregulated in presymptomatic stages of ALS in laser-capture microdissected motor neurons from the murine spinal cord ($p = 0.002$)[261]. Slc11a2 is lowly expressed in the day 3 $SOD1^{G93A}$ ESMNs of the cell autonomous condition, but by day 7 (and in the nodes connecting the day 3 with the day 7 population) has comparable expression in both genotypes. This suggests either a delayed activation of Slc11a2 in $SOD1^{G93A}$ ESMNs, or degradation of Slc11a2 mRNA in those cells. As I will talk about in the next paragraph, the transcripts for Tfrc and Slc11a2 are post-transcriptionally regulated and are tightly correlated with iron availability in the cell.

Ferric iron is converted into ferrous iron through iron reductases such as Cxcl12 and Frrs1l, which are expressed throughout the timecourse. Intracellular ferrous iron concentrations are sensed and maintained by two iron responsive proteins, Irp1 and Irp2, through post-transcriptional modifications of mRNAs encoding iron storage and transport proteins.[262] When iron availability is low, Irp2 stabilizes transcripts with a 3' iron response element (IRE) in the pre-mRNA (eg TfR1 and Slc11a2, but not Slc39a14), and destabilizes transcripts with a 5' IRE (eg Fth1, Ftl). A mouse model with Irp2 knockout leads to neurodegeneration, and specifically an ALS-like phenotype.[263, 264] This signifies that despite accumulation of ferric iron in ferritin, a non-metabolically available iron storage system, neurons can be effectively starved for metabolically active ferrous iron.

Iron response is tightly correlated with immune and stress responses such as anoxia. However, canonical transcriptional changes for these pathways (such as C reactive protein overexpression) are not seen in these data, suggesting an $SOD1^{G93A}$ specific response. In order to see if Irp2 could play a role in $SOD1^{G93A}$ mediated motor neuron degeneration, the $SOD1^{G93A}$ and $SOD1^{WT}$ mouse spinal cords were stained for Irp2 expression. While published data suggest that Irp2 is a cytoplasmic protein, strong nuclear Irp2 signal was seen in motor neurons in the cervical, thoracic, and lumbar regions of the $SOD1^{WT}$ mouse spinal cord. The nuclear localization was disrupted in the $SOD1^{G93A}$ motor neurons, and was replaced by cytoplasmic aggregation that

was p62 negative (Figure 13). This suggests altered activity by this key regulator of iron homeostasis.

Topological Representations of Days 3, 7, and 14 of the Autonomous Condition

The day 3, 7, and 14 combined data were also visualized through scTDA[74]. $SOD1^{WT}$ and $SOD1^{G93A}$ ESMNs plated over $SOD1^{WT}$ astrocytes were examined using Jaccard mutual information (Figure 14). Mutual information offers the ability to detect nonlinear correlation within transcriptional signatures, offering exquisite discrimination within highly similar ensembles of neurons[231]. Continuity between the transcriptional profiles of ESMNs sampled at day 3 and 7 was retained, and ESMNs from the two genotypes separated at day 14 into two distinct paths. The possible early aging of day 3 and 7 $SOD1^{G93A}$ ESMNs was evident as overlap between day 7 $SOD1^{WT}$ and day 3 $SOD1^{G93A}$ nodes, and day 14 $SOD1^{WT}$ and day 7 $SOD1^{G93A}$ nodes, when nodes were colored by library (Figure 14). Genes involved in two key pathways, calcium regulation and lipogenesis, were among those differentially regulated between genotypes at day 14 and are partially responsible for the bifurcation between them (Figure 15, Supplemental Figure 22C).

Single nuclei sequencing of the $SOD1^{G93A}$ spinal cord

Isolating cells from the spinal cord is challenging because dissociation is inhibited by the high myelin content in the tissue. Tissue homogenization followed by nuclei purification through centrifugation is an alternate method of gaining access to transcriptional information in single cells. As a proof of concept, single nuclei from a $SOD1^{G93A}$ end-stage mouse were isolated and sequenced (Figure 16A,B,C). Glial and neuronal nuclei showed discrete transcriptional programs (Figure 16D). Subpopulations of neurons (excitatory and inhibitory interneurons) as well as highly stressed cells could be identified via clustering (Figure 16E). Glial populations (astrocytes, microglia, oligodendrocytes) and endothelial cells were also identified (Figure 16F). This pilot experiment shows that single nuclei sequencing can discern differences in cell populations and be used as a platform to study *in vivo* ALS disease progression with high resolution.

Methods

Cell culture and single cell isolation

ESMNs were differentiated according to published protocols, and co-cultured on poly-D lysine and laminin coated coverslips with primary astrocytes harvested from p2 mice.[34, 65] At days 3, 7, and 14, cells were dissociated off coverslips using the Worthington papain digest system (Worthington cat. # PDS) with >95% viability. Single GFP+ DAPI- ESMNs were sorted into a 384 well plate containing 1ul of PBS+1:100 superasin. 384 well plates were immediately frozen in liquid nitrogen and stored at -80C before processing.

Spinal cord nuclei purification and isolation

The lumbar region of the spinal cord was grossly dissected from a sacrificed p150 $SOD1^{G93A}$ mouse in one minute to preclude overwhelming transcription of stress genes. The spinal cord was immediately placed into ice cold homogenization buffer (1x salt solution [5mM $CaCl_2$, 3mM $Mg(Ac)_2$, 10mM Tris HCl pH 7.5], 1mM b-mercaptoethanol, 320mM sucrose, 0.1mM EDTA, 0.1% NP-40) with DAPI. Homogenization was performed 25 times with a lose pestle followed by 20 times with a tight pestle on ice. The homogenized solution was filtered through a 100um mesh and incubated with primary Cy3-conjugated NeuN antibody (abcam ab104225, ab188287) for 10 minutes on ice. An equal volume of 50% optiprep solution (diluted in 1x salt solution, 1mM b-mercaptoethanol, 320mM sucrose) was added, mixed gently, and layered over 10ml of 29% optiprep solution (diluted in 1x salt solution, 1mM b-mercaptoethanol, 320mM sucrose). The gradient was spun at 10,100g in a swinging bucket ultracentrifuge for 30mins at 4C. The upper layer was fully removed first to prevent carryover of debris, then the bottom layer was also fully removed. Pelleted nuclei were resuspending in PBS with 1:100 superasin. Neuronal (Cy3+) and glial (Cy3-) nuclei were sorted into separate 384 well plates.

Single cell library generation

Library generation was done using a modified version of SCRB-seq, with 4 plates done in parallel.[265] Briefly, 1ul of a 2uM reverse transcription primer containing a universal primer, cell specific barcode, and UMI was added to each well. Plates were incubated at 72C and placed immediately on ice. 3ul of RT master mix (864ul 10mM dNTP, 1728ul 5x RT buffer, 69ul 1:5*10^6 ERCC, 175ul RNase inhibitor, 150ul Maxima H-, 2200ul water) were added to each well, and RT was performed with 42C extension for 90 mins, followed by 10 cycles of 50C 2mins and 42C 2 mins. The reaction was terminated with a 15 min 70C step. 7ul of a PCR master mix (40ul 100uM PCR primer, 4ml 5x Kapa HiFi Buffer, 600ul 10mM dNTP, 400ul KAPA HiFi Polymerase, 7ml water) was added per well, and PCR was done with 98C 3 mins, 15 cycles of 98C 15sec, 67C 30sec, 72C 6 mins, followed by a 72C 5 min extension. All wells from a single plate were pooled and 0.8x ampure was performed. cDNA was eluted in a 60ul volume of water. Tagmentation was done on cDNA from each plate according to Illumina Nextera XT protocol. Tagmented cDNA was amplified using a unique N7XX index primer and universal P5 primer with a 72C extension step for 3mins, followed by 95C for 30 sec and 12 cycles of 95C 10sec, 55C 30sec, 72C 1min, followed by a 72C extension for 5min. Libraries were purified by one round of 0.8x ampure followed by one round of 0.65x ampure with a final elution volume of 20ul in water. Libraries were KAPA quantified and submitted for sequencing on a NovaSeq.

Immunofluorescence

$SOD1^{WT}$ and $SOD1^{G93A}$ animals were perfused with PBS followed by 4% PFA. Spinal cords were dissected and allowed an overnight post-fixation at 4C, then cryopreserved in 30% sucrose, sectioned into cervical, thoracic, and lumbar segments, and embedded in OCT. IRP1 antibody (1:200 Santa Cruz #sc-166022) and IRP2 antibody (1:500 Novus biological #NB100-1798) were incubated on sections in blocking buffer (5% donkey serum, 1% BSA, 0.3% Triton X-100) overnight at 4C, washed, and secondary labeled with abcam Alexa-fluor conjugated pre-adsorbed antibodies. Imaging was performed on a Axio Imager Z2 with an Andor iXon3 EMCCD camera.

Processing of scRNA-seq Data

Reads were demultiplexed and deduplicated using UMItools[162]. Processed reads were mapped against a combined mm10 and ERCC reference genome using STAR mapper[166]. Cells were filtered based on processed read counts (cells with >40,000 reads and <1,000,000 reads were retained) and percent ERCC (cells for which ERCC counts accounted for >0.15% of reads were discarded). Counts for genes detected per cell were normalized to transcript per million (TPM). kNN clustering and force-directed graphing was done using SPRING[240]. scTDA visualization was done using the Ayasdi platform. Spinal nuclei clustering was done using RaceID[77].

Future Directions

Topologically mapping single cells in a disease timecourse has not yet been done, and topological modeling of this data poses new computational challenges. Combining 4 separate timecourse studies into a single representation is an attractive approach, and a platform for intersecting these data is being developed. Computational analysis will require new statistical frameworks to identify pseudotime in the representation.

The role of iron homeostasis in ALS disease progression will be examined more closely. Ferric (Prussian Blue) and ferrous (Turnbull Blue) iron stains can reveal the relative concentrations of stored and available iron in spinal cord tissue. While ferric iron buildup is a common symptom of neurodegenerative disorders, and ferric iron chelators are being examined as possible therapeutic tools in ALS[266], the relative levels of metabolically available ferrous iron should be considered. Furthermore, the activity of Irp1 and Irp2 should also be measured both in motor neurons and surrounding glial cells. This can be done by looking at the mRNA levels of transcriptional targets containing IRE. A comparative study of transcriptional programs in nuclei isolated from the $SOD1^{WT}$ and $SOD1^{G93A}$ spinal cords should be done to assess the differential expression of IRE containing transcripts *in vivo*.

While *in vitro* human models of ALS do not show selective MN degeneration like the ESMN models do, they do exhibit aberrant RNA and protein aggregation, transcriptional changes, and axonal breakage. The lack of degeneration phenotype could be a result of the late-onset of ALS, with the newly generated iMNs not aging sufficiently *in vitro* to exhibit end-stage properties. Alternatively, iMNs may be sufficiently artificial that they cannot respond to the presence of a disease-associated gene in the same manner as purified ESMNs. Purity or heterogeneity in iMNs and hES-MNs has not been well characterized. One approach to purifying iMNs is by introducing a reporter gene, such as is done with Hb9::eGFP in mESC-derived models. While this can be achieved with human stem cells, the characteristics of differentiation may be different and the resulting population may have more variability or artificiality than in mESC-derived models. Using scRNA-seq to characterize the heterogeneity in iMNs and hES-MNs prior to disease modeling can help interpret biological results from these studies.

Discussion

Single cell RNA sequencing studies provide us with the unique capability of measuring the transcriptional activity of individuals within a heterogeneous population. This approach allows the determination of cellular subpopulations, and, coupled with timecourse studies, altered transcriptional dynamics and pathway activation. In order to gain access to this information, computational approaches with sufficient sensitivity and statistical strength need to implemented. Single cell topological data analysis (scTDA[74]) is a powerful approach that enables in-depth studies of transcriptional progression, and in this thesis was applied to *in vitro* neuronal differentiation and ALS disease progression.

Early transcriptional alterations in ALS have been difficult to identify[261]. However, given that these events may mediate the onset of ALS and catalyze disease progression, understanding them is of critical importance. Due to the asynchronous progression of ALS in subsets of motor neurons within the spinal cord, bulk sequencing studies of early transcriptional changes are easily drowned out by variability of expression within cells and the low representation of susceptible motor neurons in tissue. *in vitro* modeling allows for a purified population of motor neurons to be studied, thus enriching for the target cell type. Furthermore, the experimental setup allows for the dissection of cell autonomous changes, driven by the presence of the $SOD1^{G93A}$ mutation within motor neurons themselves, from non-cell autonomous changes, driven by the presence of the $SOD1^{G93A}$ mutation in co-cultured astrocytes. In order to understand the relative contributions of these pathways, I performed scRNA-seq across 3 timepoints (days 3, 7, and 14) on mESC-derived motor neurons (ESMNs) overexpressing either $SOD1^{WT}$ or $SOD1^{G93A}$ cultured over primary murine astrocytes overexpressing either $SOD1^{WT}$ or $SOD1^{G93A}$. I then generated 3 distinct scTDA models to compare early transcriptional changes associated with cell autonomous (CA), non-cell autonomous (NCA), and disease culture conditions.

Visualization of the scTDA models shows unique patterns in the organization of ESMNs within the representations, which reflect differences in global transcriptional programs over time under CA, NCA, and disease culture conditions. In CA, ESMNs sampled at day 3 have similar global transcriptional profiles regardless of their genotype, as can be seen by shared and closely linked nodes. However, over time in culture, the $SOD1^{WT}$ and $SOD1^{G93A}$ ESMN populations diverge, suggesting that transcriptional alterations associated with $SOD1^{G93A}$ expression in ESMNs accumulate over time and progressively impact neuronal health. In NCA, the timecourse trajectories intersect such that the ESMNs cultured over $SOD1^{G93A}$ astrocytes sampled at day 3 reside in nodes located between days 3 and 7 of the control ESMNs. The localization of these cells is reminiscent of a (theoretical) day 5 control sample, and suggests an accelerated aging of the ESMNs driven by expression of $SOD1^{G93A}$ in astrocytes. Comparing control samples with disease samples shows convergence of ESMN populations over time. Given that by day 7 many of the vulnerable ESMNs cultured in disease conditions have degenerated, this pattern suggests that the resistant population of ESMNs is transcriptionally more similar to the control ESMNs than the vulnerable population is. These kinetics indicate that the early timepoint provides crucial information for transcriptional profiling in ALS.

Using the statistical analyses built into scTDA (centroid, dispersion, and connectivity), I identified altered gene activity patterns that defined the progression of ALS in ESMNs under CA, NCA, and disease stress. In this discussion I will focus on two pathways, iron import and oxidative-stress mediated accelerated aging, which converge on the unfolded protein response pathway through ER stress. While both iron dysregulation[267] and oxidative stress[254] have been associated in ALS, the resolution to study these phenomenon at the transcriptional level has previously been missing. Furthermore, the activation of these pathways suggest a novel mechanism of instigation of ER stress and the terminal unfolded protein response, which has been implicated as a key pathogenic feature of ALS[268].

Iron homeostasis is critical for neuronal function, and its dysregulation has been implicated in many neurodegenerative diseases[269]. Iron import is largely controlled by two proteins, transferrin receptor (Tfrc) and divalent metal transporter 1 (DMT1)[269], the expression of which is tightly regulated at the transcriptional level[270, 271]. Both of these proteins are downregulated in SOD1^{G93A} ESMNs at day 3, suggesting a kinetic delay in the activation of iron response genes. While iron accumulation has traditionally been thought to be the predominant pathogenic feature in neurodegeneration, work done on an iron response protein (IRP2) knockout mouse line has shown that effective iron deficiency results in a model that phenocopies ALS[264]. This IRP2 knockout mouse displays compromised mitochondrial function, which is an early feature of ALS and a major source of oxidative and ER stress in motor neurons[272, 273]. Impaired iron import early in SOD1^{G93A} motor neurons may poise these cells to be selectively vulnerable to oxidative damage and ER stress later in the disease, despite the recovery of transcript expression over culture time.

Mitochondrial dysfunction, oxidative stress, and ER damage are also all functions of aging, which is a major risk factor for the development of many neurodegenerative diseases including ALS[274]. The intermediate localization of day 3 NCA-stress ESMNs between the day 3 and day 7 control ESMNs in the NCA scTDA representation raises the possibility that ESMN aging could be accelerated by the presence of SOD1^{G93A} astrocytes. While aging is difficult to define transcriptionally, recent work has identified the antioxidant inhibitor Thioredoxin Interacting Protein (Txnip) as consistently upregulated in aged brains[249]. Overexpression of Txnip has also been shown to be sufficient to induce aging in a Drosophila model through decreased resistance to oxidative damage[250]. Consistent with these findings, we see elevated expression of Txnip in the day 3 NCA ESMNs. The role of Txnip in the ALS spinal cord has not been studied, however a number of studies suggest that it may play a role in both selective motor neuron degeneration and ER-stress mediated neuroinflammation.

One function of Txnip is to bind to and inhibit Thioredoxin (cytoplasmic Trx1, and mitochondrial Trx2), a vital antioxidant protein known to be involved in mitigating oxidative damage[275]. Trx1 expression is tissue specific, and has been shown to be expressed in motor neurons of the spinal cord[276]. Trx1 is upregulated after nerve injury[277] and in lesioned cortical tissue[278], and is thought to contribute to regeneration in the CNS[278]. In *in vitro* model systems, it is also secreted by U251 astrocytoma cells and promotes survival of primary murine neurons[279]. Trx1 shows the highest upregulation in lumbar spinal cord of ALS patients compared to controls (600%), where it possibly constitutes an endogenous defense mechanism in motor neurons against oxidative damage[280]. Elevated expression of Txnip inhibits the protective function of Txr by directly binding to it and sequestering it away from target proteins[281]. Thus, upregulation of Txnip in ESMNs may leave them poised for susceptibility to oxidative stress.

Txnip is upregulated through Atf5, a downstream effector of ER stress and the unfolded protein response (UPR)[282]. In type 1 and type 2 diabetes, this Txnip activation pathway has been shown to switch cells from a protective ER response to a terminal apoptotic UPR through the activation of the NLRP3 inflammasome[282, 283]. Atf5 is also upregulated in NCA day 3 ESMNs, and may explain the upregulation of Txnip. Through this mechanism, the ER stress response in ESMNs may play a role in NCA ALS-associated motor neuron generation.

Conclusion

The mammalian spinal cord conveys complex somatosensory and motor signals, the propagation of which depends on the interactions of a diverse ensemble of neurons and glia. During amyotrophic lateral sclerosis (ALS) disease progression, altered function amongst all of these cells contributes to selective motor neuron death. In this thesis, I have outlined single cell transcriptomic and computational approaches to dissect the relative contribution of these cell types to disease progression.

To do so, I have optimized single cell and single nuclei sequencing approaches of *in vitro* and *in vivo* models to be able to determine transcriptional definitions corresponding to cellular state, with high depth of coverage. Contending with continuous structure in an unsupervised manner within single cell sequencing data describing heterogeneous cellular responses was previously computationally lacking in feasibility. As described, Topological Data Analysis enables the direct determination of transcriptional programs accompanying differentiating ensembles of single cells[74]. The work presented in this thesis leaves the field well poised to determine cellular markers associated with transitory cell states and their associated patterns of gene regulation, and was applied to the study of cell autonomous and non-cell autonomous contributions to ALS disease progression.

Furthermore, this work comprises an important technological platform to delineate the transcriptional responses triggered by inductive cues. A defined transcriptional program (such as one responsible for neurogenic commitment of a neural progenitor cell, or motor neuron cell death in ALS) can be initiated or disregarded in individual cells. Whether or not it is enacted depends both on the identity and molecular properties of the cell, and the stochastic nature of gene expression[73]. A computational approach for determining to what extent underlying heterogeneity is responsible for differences in transcriptional pathway activation, or to what extent the transcriptional activation is a result of stochastic choice in a clonal population, remains to be

developed. Especially in context of ALS, where disease onset occurs in a stochastic location[284], understanding the genetic landscape that promotes pathology is critical. Using the high resolution scRNA-seq timecourse experiments presented here, I hope to be able to better understand the transcriptional settings that steer cells towards developing disease phenotypes or surviving them.

The single cell sequencing field is progressing rapidly. On the leading edge are multi-modal investigations where simultaneous readout of seemingly disparate biological observables are measured from individual cells. By way of example, protein abundance and RNA content, chromatin accessibility and RNA content, and methylation are being integrated to expand depth of coverage and mechanistic insight into the origins of diversity and heterogeneous cellular responses that accompany[285, 286]. Furthermore, spatially resolved transcriptomic approaches offer insights into cellular individuality, while simultaneously revealing a Cartesian coordinate axis associated with cellular diversity and response[287]. All of these technological directions serve a valuable function beyond validation of single cell measurements. They confer upon the experimentalist the ability to develop approaches to interrogate cell type specific interactions and exchange of information. This, of course, is of paramount importance when considering cell autonomous and non-cell autonomous changes, such as those associated with localized neurodegenerative events and changes in glia-neuron communication. Furthermore, spatially resolved approaches offer much to the understanding of the spread of ALS. It is well documented that the pathology progresses laterally and rostro-caudally from the location of onset, however the mechanism underlying disease spread is poorly understood[288, 289].

Taken together, the future is bright for single cell technologies and their capacity to yield insights into ALS disease progression.

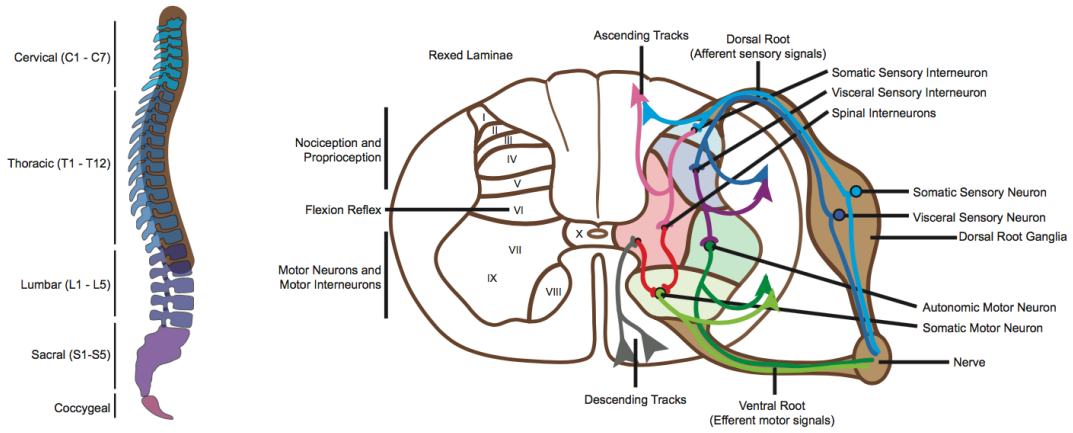


Figure 1. Schematic representation of neuronal organization in the spinal cord.

Motor neurons are located in the ventral horn (rexed laminae vii, viii, ix).

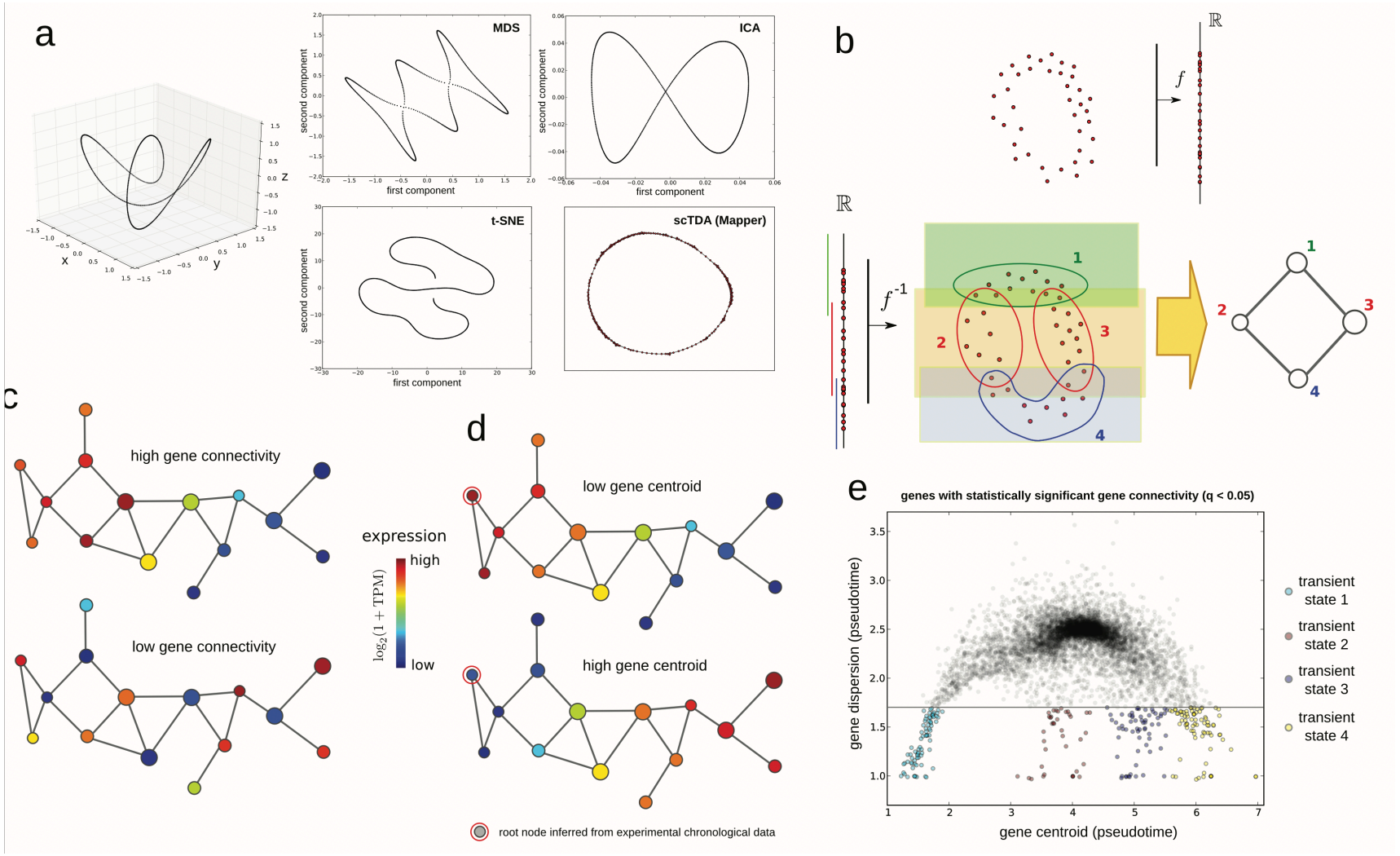

a
MDS
ICA
t-SNE
scTDA (Mapper)
first component
second component
x
y
z
b
$\mathbb{R}$
f
f^{-1}
1
2
3
4
c
high gene connectivity
low gene connectivity
expression
high
low
$\log_2(1+\mathrm{TPM})$
d
low gene centroid
high gene centroid
root node inferred from experimental chronological data
e
genes with statistically significant gene connectivity (q < 0.05)
gene dispersion (pseudotime)
gene centroid (pseudotime)
transient state 1
transient state 2
transient state 3
transient state 4

Figure 2. Topological analysis of longitudinal single-cell RNA-seq data.

a. Comparison of several methods for reducing the dimensionality of data. A toy example is shown, illustrating the artifacts that can emerge when standard dimensional reduction methods are used to represent differentiation trajectories. A total of 1,000 points are sampled from a twisted circle in three-dimensional space. MDS, ICA, t-SNE, and Mapper were utilized to represent the above points in two dimensions. Of these methods, only Mapper was able to capture the continuous circular trajectory of the three-dimensional space without introducing artificial intersections or disrupting the trajectory.

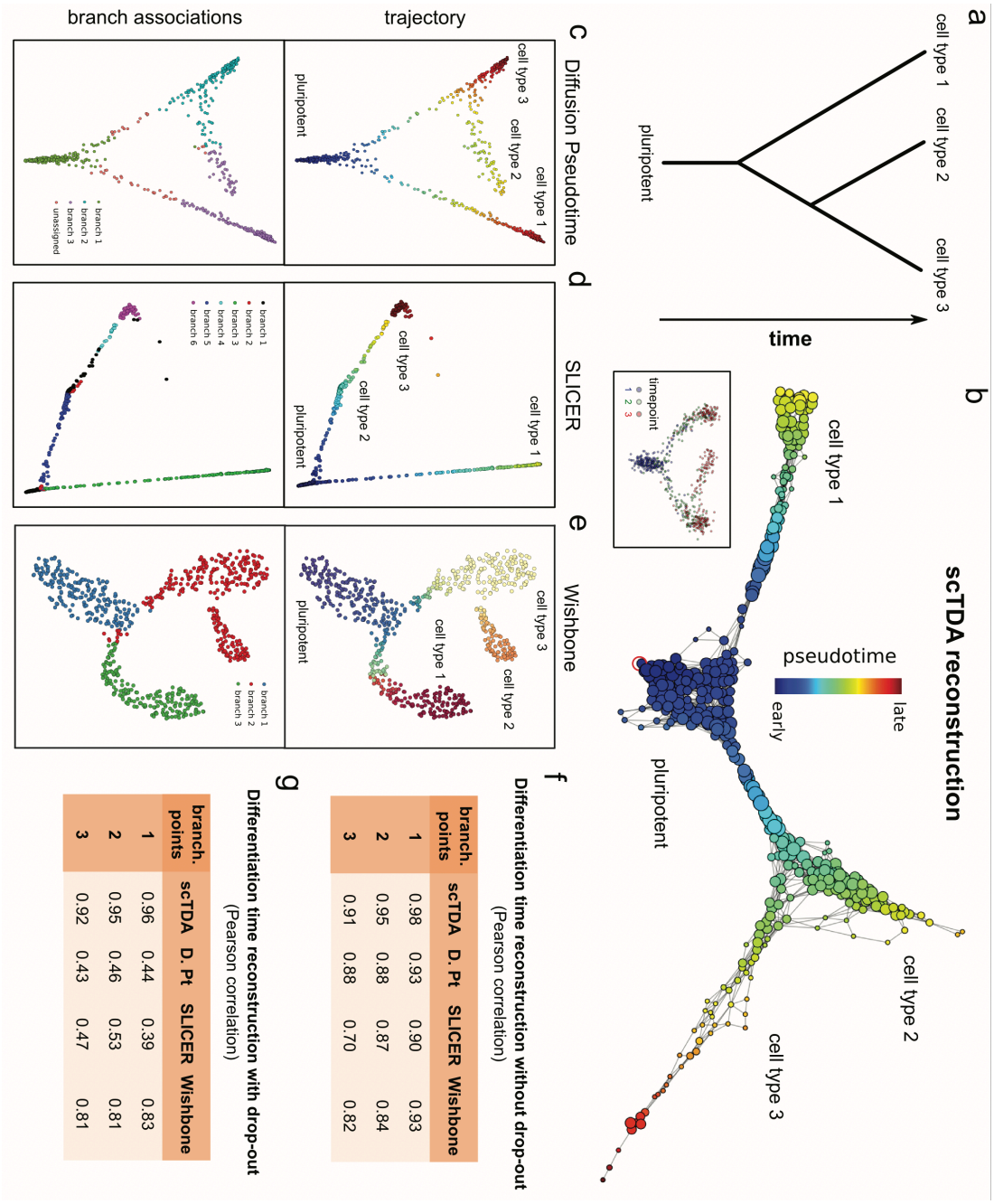
b. A schematic illustrating the Mapper algorithm. Top: Mapper builds upon any dimensional reduction function *f* mapping the high-dimensional single-cell RNA-seq point cloud data into $\mathbb{R}^k$ (for simplicity we take $k = 1$ in this figure). Bottom: under the inverse function f^{-1}, a covering of $\mathbb{R}^k$ maps into a covering of the single-cell point cloud data. Clustering is performed independently in each of the induced patches in the high-dimensional space. In the low-dimensional representation, a node is assigned to each cluster of cells. If two clusters intersect, the corresponding nodes are connected by an edge. Topological features in the low dimensional representation are guaranteed to also be present in the original high-dimensional RNA-seq space.

c. Gene connectivity. Gene connectivity allows for the identification of genes that are differentially expressed by a cellular subpopulation of the differentiation process, without predefining any cellular subpopulation. Represented is a toy example of two genes with very different gene connectivity on the topological representation. Top: An example of a gene with high gene connectivity in the topological representation. This signifies that there is a set of cells with similar global expression profiles and high expression levels of the gene. Bottom: An example of a gene with low gene connectivity in the topological representation.

d. Illustration of gene centroid. The centroid of a gene, measured in pseudotime, quantifies where the expression of a gene sits in the topological representation with respect to the root node. The root node represents the least differentiated cellular state, and is determined from the experimental sampling times. A toy example of two genes with very different centroid can be used

to illustrate the concept. Top: a gene with low value for the expression centroid, being mostly associated to pluripotent cells. Bottom: a gene with a high value for the centroid, being mostly associated to differentiated cells.

e. Identification of transient cellular states. Transient cellular states are identified in an unsupervised manner by clustering low-dispersion genes with significant gene connectivity according to their centroid in the topological representation. In the figure, an example of distribution of centroids and dispersions for genes with significant gene connectivity is shown. Four clusters of low-dispersion genes are identified, which correspond to four transient cellular states arising throughout the differentiation process.



branch. points	scTDA	D. Pt	SLICER	Wishbone
1	0.98	0.93	0.90	0.93
2	0.95	0.88	0.87	0.84
3	0.91	0.88	0.70	0.82

branch. points	scTDA	D. Pt	SLICER	Wishbone
1	0.96	0.44	0.39	0.83
2	0.95	0.46	0.53	0.81
3	0.92	0.43	0.47	0.81

Figure 3. Comparison of several algorithms for ordering cellular states.

a. A noisy, branched, asynchronous cellular differentiation process was simulated. The differentiation tree of the process is represented. 700 cells were sampled from this process at three time points. Using this data, we attempted to reconstruct the structure of the differentiation tree with scTDA, alongside the algorithms Diffusion Pseudotime, SLICER, and Wishbone, which rely on branch assignments for downstream statistical analysis.

b. Reconstructed differentiation trajectory using scTDA. scTDA recovered the structure of the simulated differentiation process and correctly rooted the tree using the experimental chronological information. Nodes correspond to sets of cells sharing similar global transcriptional profiles, with the node sizes proportional to the number of cells in the node. Nodes that are connected by an edge have at least one cell in common. For reference, an inset with the latent MDS representation of the data, colored by sampling day, is also shown.

c. Reconstructed differentiation trajectory using Diffusion Pseudotime. While the representation of the data using the first two diffusion coefficients reproduces the structure of the differentiation tree, the branches were not correctly assigned.

d. Reconstructed differentiation trajectory using SLICER. The representation constructed by SLICER using locally linear embedding was unable to capture the complete structure of the differentiation tree and branch assignments.

e. Reconstructed differentiation trajectory using Wishbone. The t-SNE representation of the data used by Wishbone reproduces the structure of the differentiation tree and identified correctly the first branching point. However, Wishbone was unable to identify the second branching point.

f, g. Pearson's correlation coefficient between the pseudo-time, inferred from the data by scTDA, Diffusion Pseudotime (D. Pt), SLICER, and Wishbone, and the actual simulated differentiation time. Cellular differentiation processes with one, two, or three branching points were considered, both in the absence (**f**) and the presence (**g**) of drop-out events.

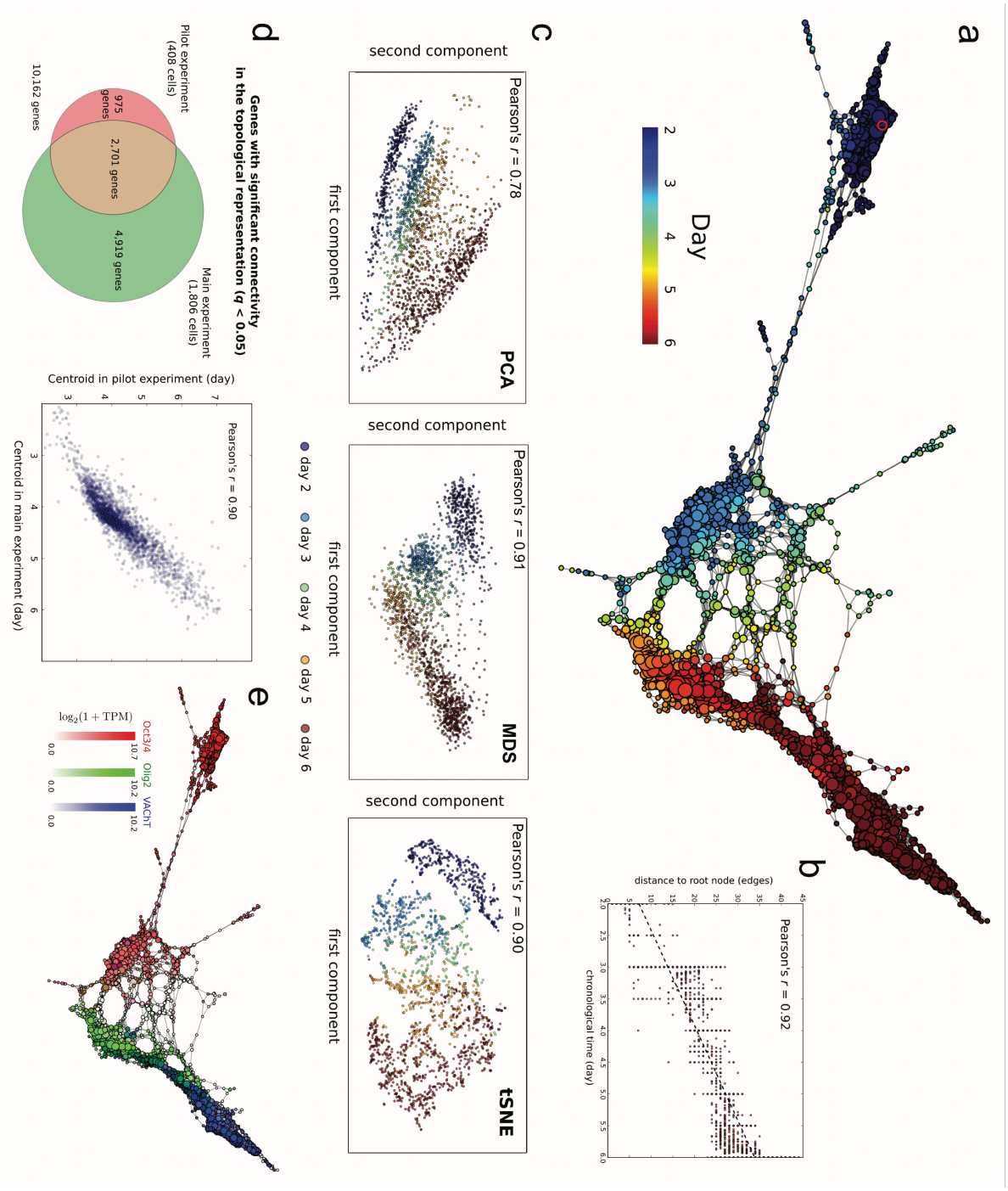


Figure 4. Topological representation of longitudinal single-cell RNA-seq data from the differentiation of mESCs into MNs.

a. Topological data analysis recapitulates chronological order based on expression data alone. The topological representation of the expression data of 1,964 single cells, sampled from the differentiation of mESCs into MNs, is labeled by sampling time. The root node, inferred from the experimental chronological information, is indicated with a red circle.

b. The distance of each node to the root node, represented as a function of sampling time. The chronological time of a node is defined as the mean of the sampling times of the cells in the node.

c. Comparison to standard dimensional reduction algorithms. Dimensional reduction of the same expression data of 1,964 single cells, sampled from the differentiation of mESCs into MNs, using PCA, MDS, and t-SNE. The Pearson's correlation coefficient between the sampling time and the two-dimensional Euclidean distance to the root cell (defined as the one that maximizes this correlation) is indicated in each case.

d. Consistency between main and pilot experiments. Left: Venn diagram of genes with significant gene connectivity ($q < 0.05$) in the topological representations of the two datasets. Both experiments are highly consistent in their calls (Fisher exact test p-value $< 10^{-100}$). The number of significant genes is larger in the main experiment, consistent with its higher statistical power (due to the larger number of cells considered). Right: correlation between the centroid (expressed in pseudo-time) of the $n = 2{,}701$ genes with significant ($q < 0.05$) gene connectivity in both topological representations. The distribution of transcript centroids is highly consistent across the two experiments.

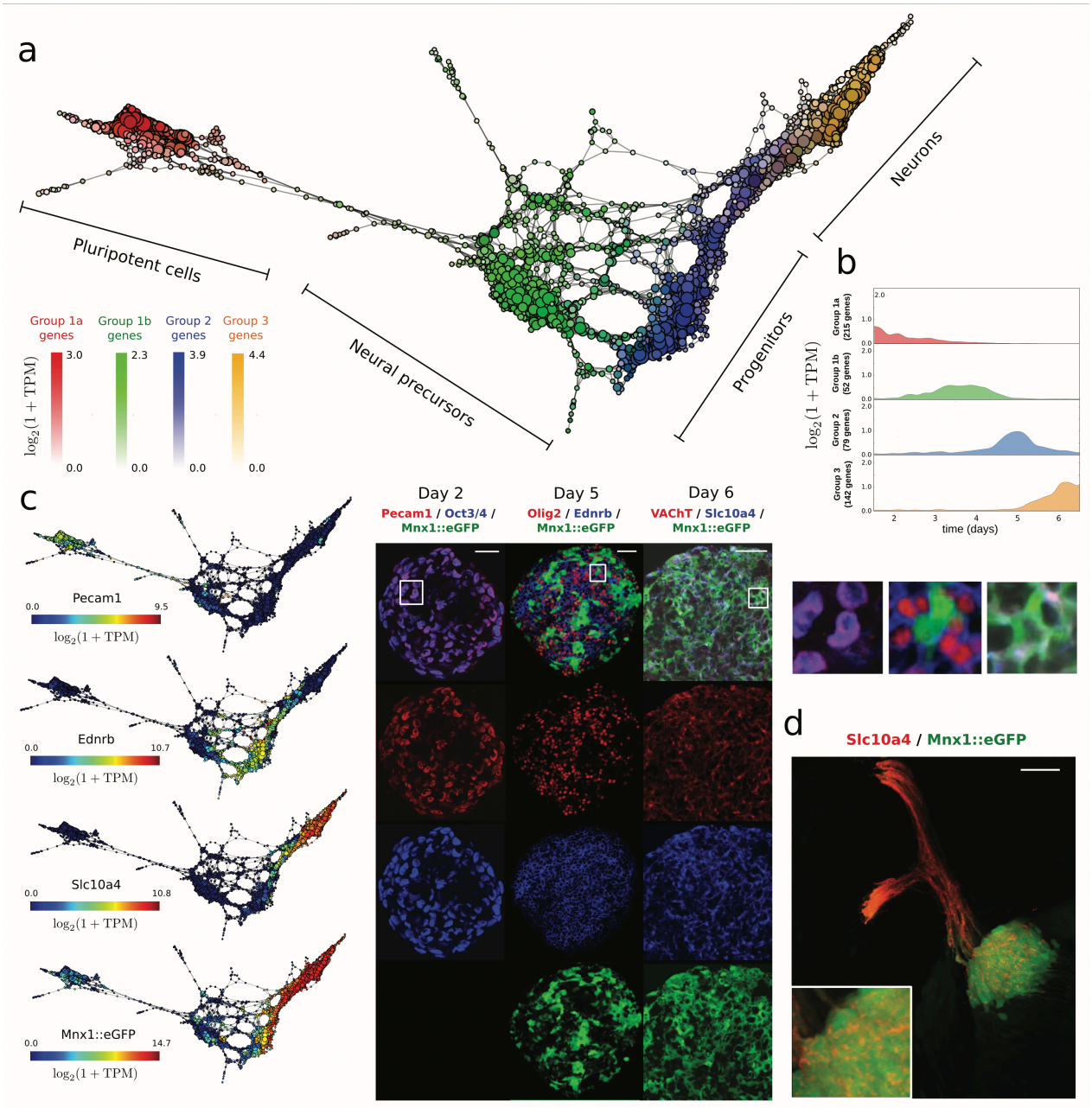


Figure 5. Cellular populations arising throughout the differentiation of mESCs into MNs and novel candidate surface markers.

a. scTDA identifies four transient populations arising throughout the differentiation of mESCs into MNs. Represented is the topological representation colored by mRNA levels of genes belonging to these four groups of low-dispersion genes, corresponding to pluripotent, precursor, progenitor

and post-mitotic populations. In total, 488 genes were uniquely assigned to these four populations based on their expression profiles in the topological representation.

b. Reconstructed expression timeline for each of the four groups of low-dispersion genes.

c. Direct detection of state specific cell surface markers identified by scTDA. Topological representation colored by mRNA levels of surface proteins *Pecam1*, *Ednrb* and *Slc10a4*, and immunostaining of cultured EBs. The scale bar in the immunostaining images denotes a length of 50 mm. Details of three regions are presented on the right. For reference, the topological representation colored by mRNA levels of the Mnx1::eGFP reporter is also shown.

d. *In vivo* validation of the motor neuron surface marker *Slc10a4*. Spinal cord section of an E9.5 mouse immunostained for *Slc10a4* (red). The pool of motor neurons is also marked by Mnx1::eGFP expression (green). The scale bar denotes a length of 50 mm.

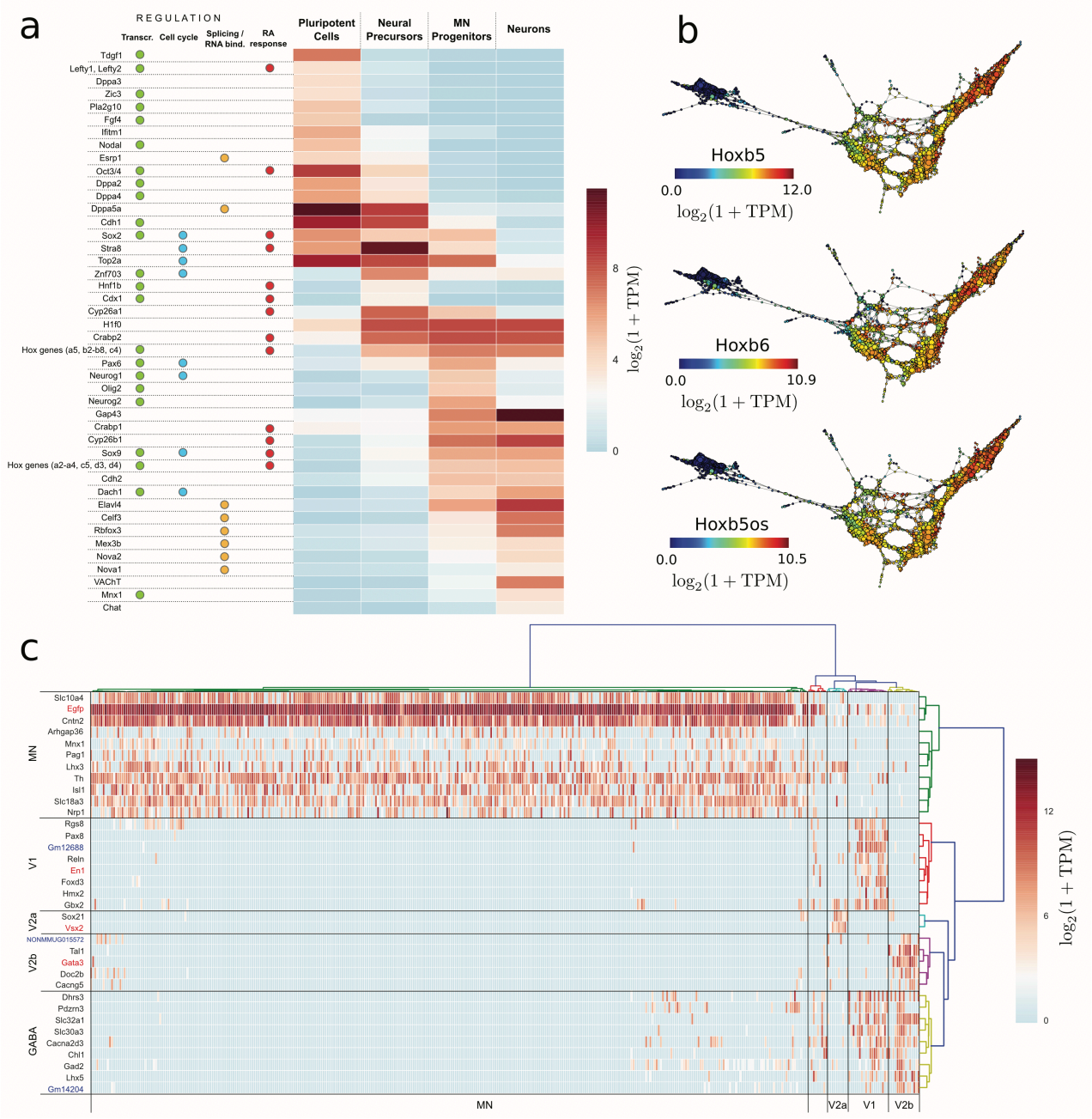


Figure 6. Differentially expressed transcripts associated with neurogenesis.

a. Differentially expressed regulators and downstream genes across the four populations arising throughout the differentiation. Genes are annotated according to their role in transcriptional, cell cycle, RNA binding protein regulation and RA response.

b. Topological representations labeled by mRNA levels of *Hoxb5*, *Hoxb6* and the antisense lncRNA *Hoxb5os*, showing concordant expression of these transcripts during the generation of motor neurons from mESCs.

c. Post-mitotic neuronal populations. Differentially expressed genes between *Vsx2*$^{+}$, *Gata3*$^{+}$ and *En1*$^{+}$ cells that were marked as post-mitotic neurons by the scTDA analysis. Hierarchical clustering of cells leads to five groups, four of which correspond to MNs, and V1, V2a and V2b interneurons. Hierarchical clustering of genes produces four groups of genes, uniquely expressed by each of the above cell types, and a fifth group associated to GABAergic neurons. lncRNAs are marked in blue.

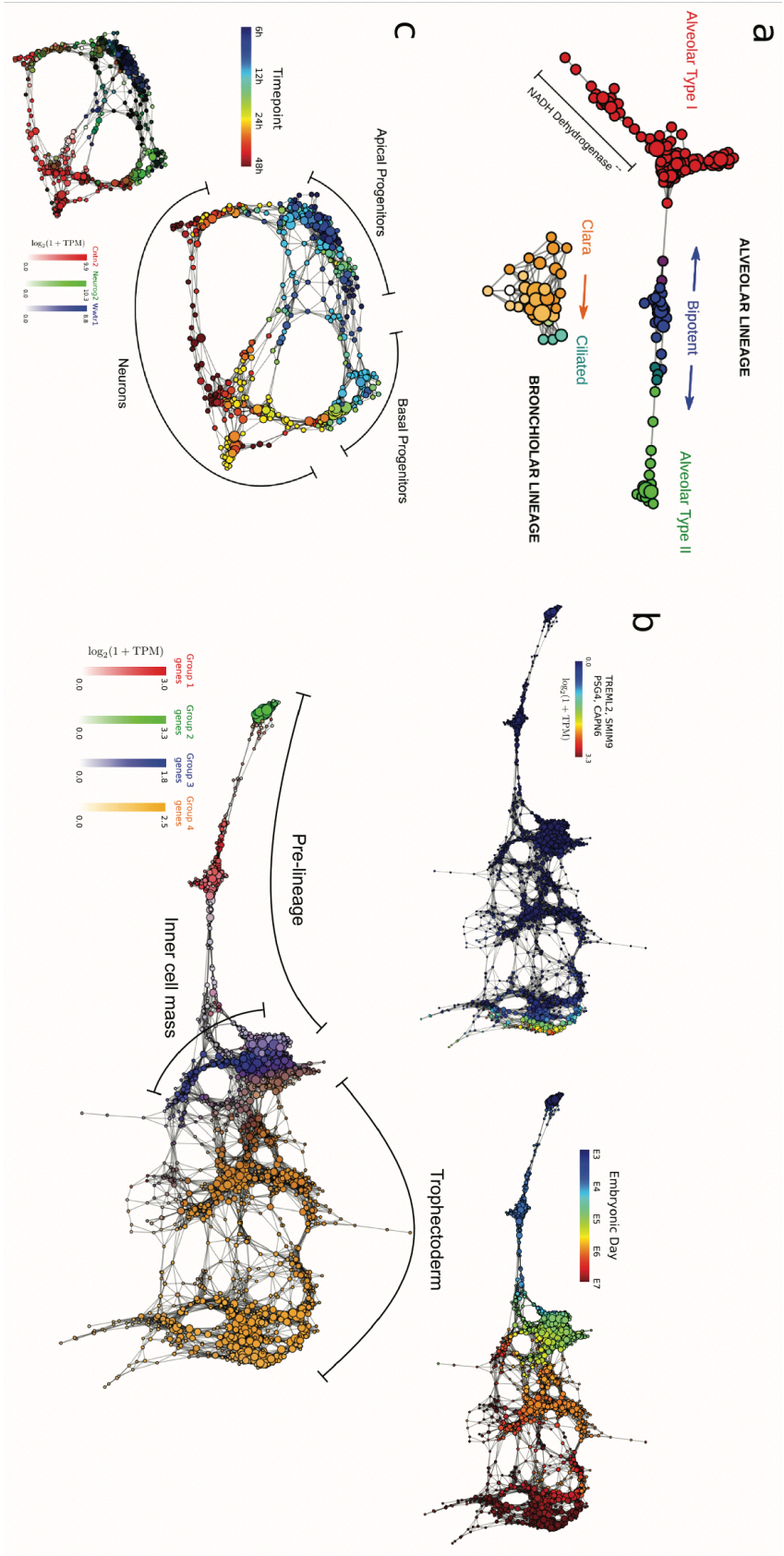

a
ALVEOLAR LINEAGE
Alveolar Type I
Bipotent
Alveolar Type II
NADH Dehydrogenase
Clara
Ciliated
BRONCHIOLAR LINEAGE
b
TREML2, SNIMM9
PSG4, CAPN6
log2(1 + TPM)
Embryonic Day
E3
E4
E5
E6
E7
Trophectoderm
Pre-lineage
Inner cell mass
log2(1 + TPM)
Group 1 genes
Group 2 genes
Group 3 genes
Group 4 genes
c
Timepoint
6h
12h
24h
48h
Apical Progenitors
Basal Progenitors
Neurons
log2(1 + TPM)
Cntn2
Neurog2
Wnt6-1

Figure 7. Application of scTDA to several *in vivo* datasets.

a. Topological representation of 80 embryonic (E18.5) mouse lung epithelial cells labeled according to cell type. scTDA correctly resolves the alveolar and bronchiolar lineages, and identifies a previously unreported set of cells with low expression of NADH dehydrogenase.

b. Topological representation of 1,529 individual cells from 88 human preimplantation embryos. Top-left, bottom: topological representation labeled by expression levels of genes associated to cellular populations arising throughout the differentiation identified by scTDA. scTDA correctly identifies, absent supervision, the segregation of the trophectoderm and inner cell mass from pre-lineage cells (bottom), as well as a polar trophectoderm (top-left). Top-right: topological representation labeled by embryonic day.

c. Topological representation of 272 newborn neurons from the mouse neocortex, labeled by sampling time after mitosis (top) and expression levels of *Cntr2*, *Neurog2*, and *Wwtr1* (bottom). scTDA correctly recapitulates converging developmental relations between apical and basal progenitors, and neurons.

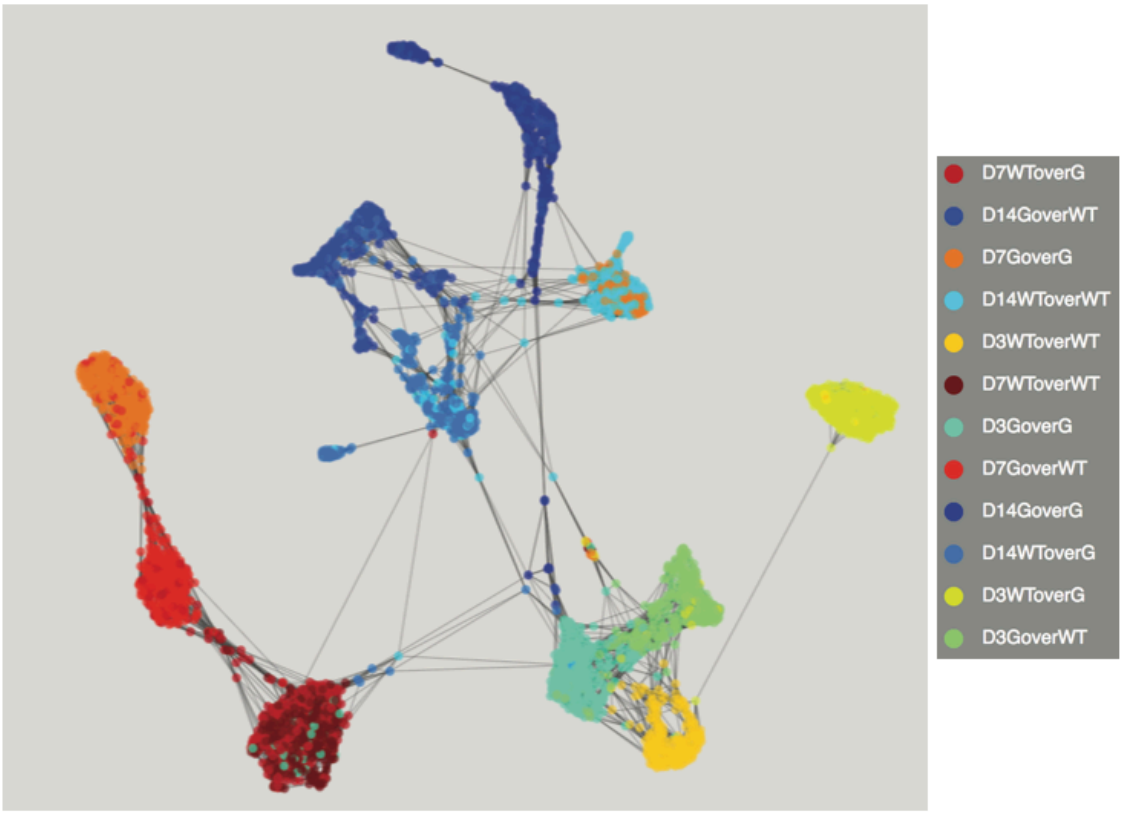


Figure 8. kNN clustering of timecourse ESMN data

Individual cells are shown colored by the condition they were sampled. Clusters of cells with similar transcriptional programs can be seen. These clusters are identified in table 1 and supplementary figure 21.

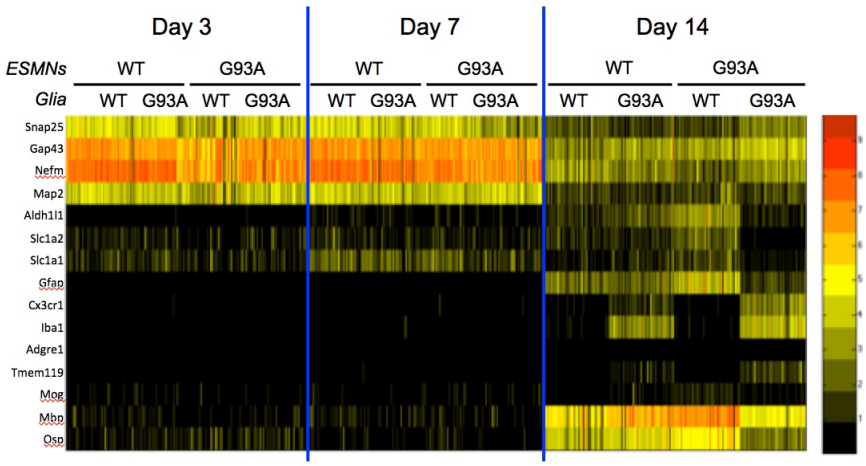
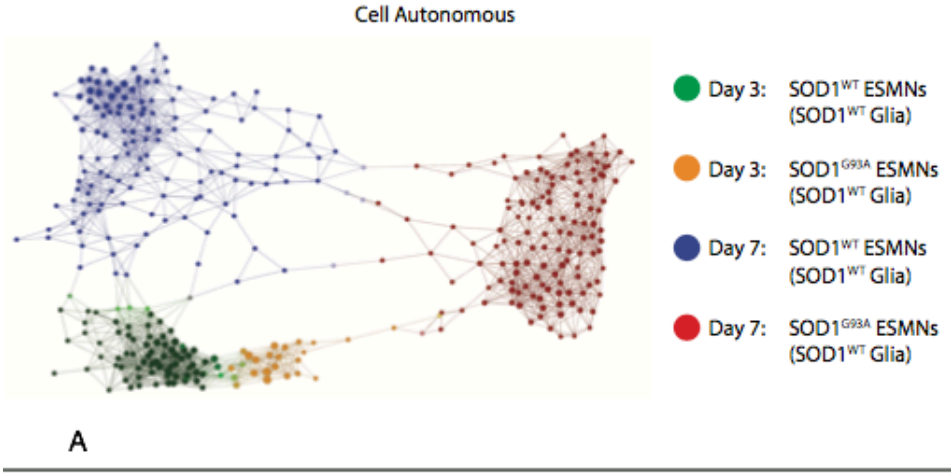
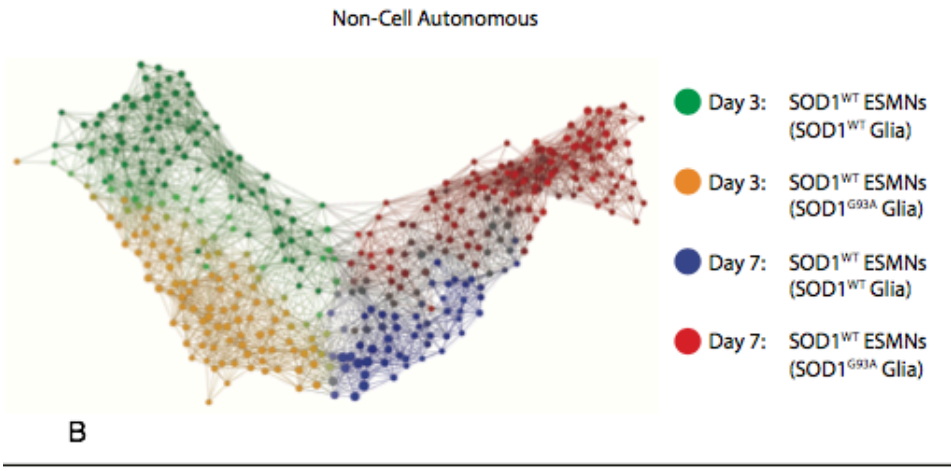


Figure 9. Heatmap of cell type specific transcriptional signatures

FACS-purified eGFP+ ESMNs show contaminating transcriptional signatures of astrocytes, microglia, and oligodendrocytes after 14 days in culture.


Cell Autonomous
Day 3: SOD1WT ESMNs (SOD1WT Glia)
Day 3: SOD1G93A ESMNs (SOD1WT Glia)
Day 7: SOD1WT ESMNs (SOD1WT Glia)
Day 7: SOD1G93A ESMNs (SOD1WT Glia)
A


Non-Cell Autonomous
Day 3: SOD1WT ESMNs (SOD1WT Glia)
Day 3: SOD1WT ESMNs (SOD1G93A Glia)
Day 7: SOD1WT ESMNs (SOD1WT Glia)
Day 7: SOD1WT ESMNs (SOD1G93A Glia)
B

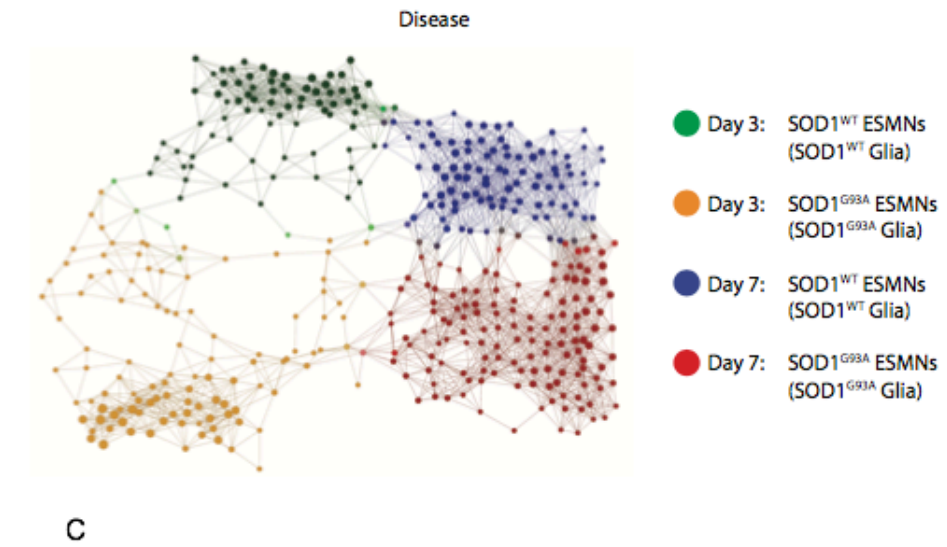

Disease
Day 3: SOD1WT ESMNs (SOD1WT Glia)
Day 3: SOD1G93A ESMNs (SOD1G93A Glia)
Day 7: SOD1WT ESMNs (SOD1WT Glia)
Day 7: SOD1G93A ESMNs (SOD1G93A Glia)
C

Figure 10. Topological representations of cell autonomous, non-cell autonomous, and disease changes in days 3 and 7 of the ESMN timecourse

A) Cell autonomous changes between $SOD1^{WT}$ and $SOD1^{G93A}$ ESMNs plated over $SOD1^{WT}$ astrocytes

B) Non-cell autonomous changes between $SOD1^{WT}$ ESMNs plated over $SOD1^{WT}$ and $SOD1^{G93A}$ astrocytes

C) Disease changes between $SOD1^{WT}$ and $SOD1^{G93A}$ ESMNs plated over $SOD1^{WT}$ and $SOD1^{G93A}$ astrocytes

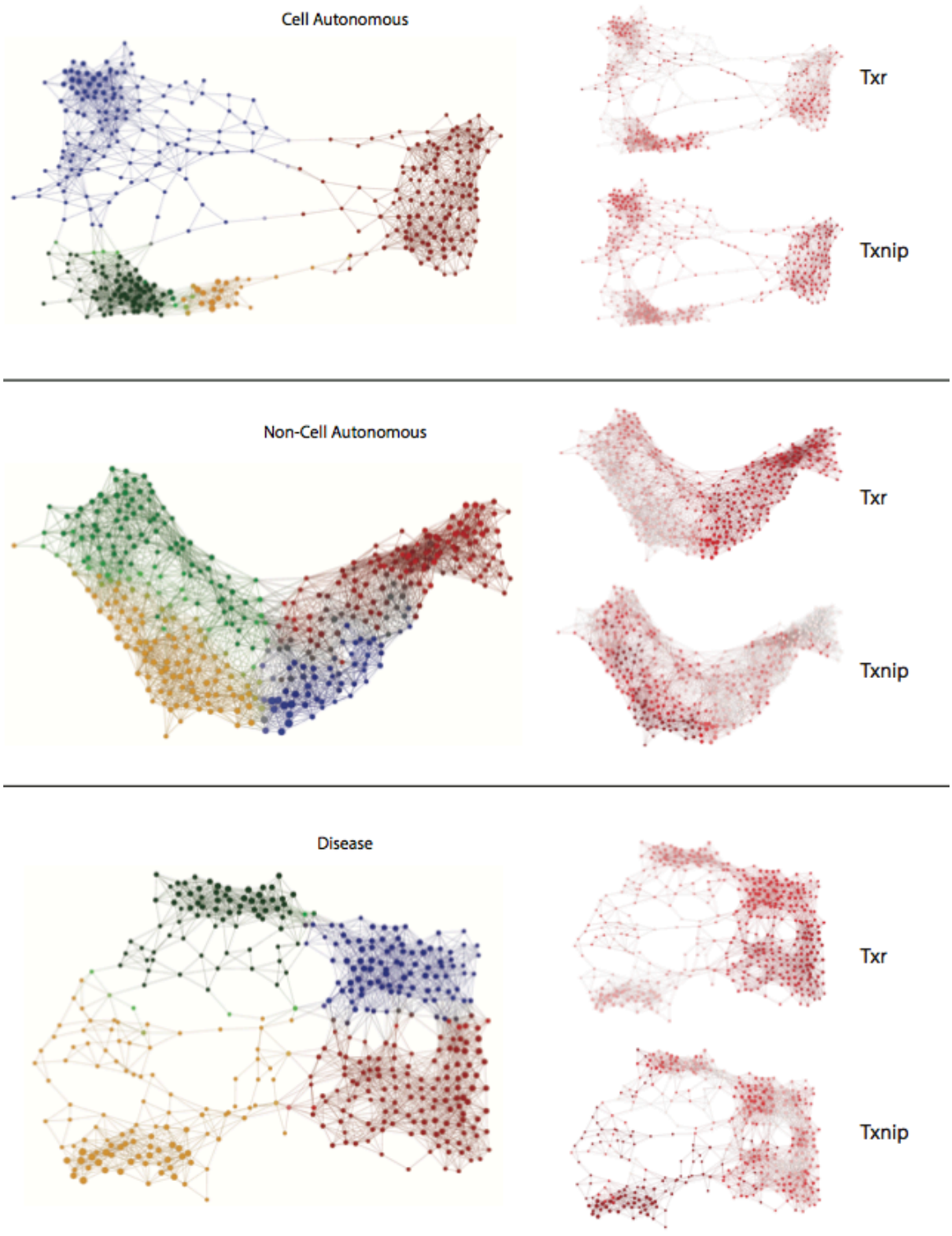


Figure 11. Changes in Txr/Txnip activation in day 3 SOD1^{WT} and SOD1^{G93A} ESMNs mediated by SOD1^{G93A} astrocytes

Relative expression of Txn1 and Txnip across conditions. (Txn1 expression log scale 10 – 12.5; Txnip expression log scale 0.5 – 2.5)

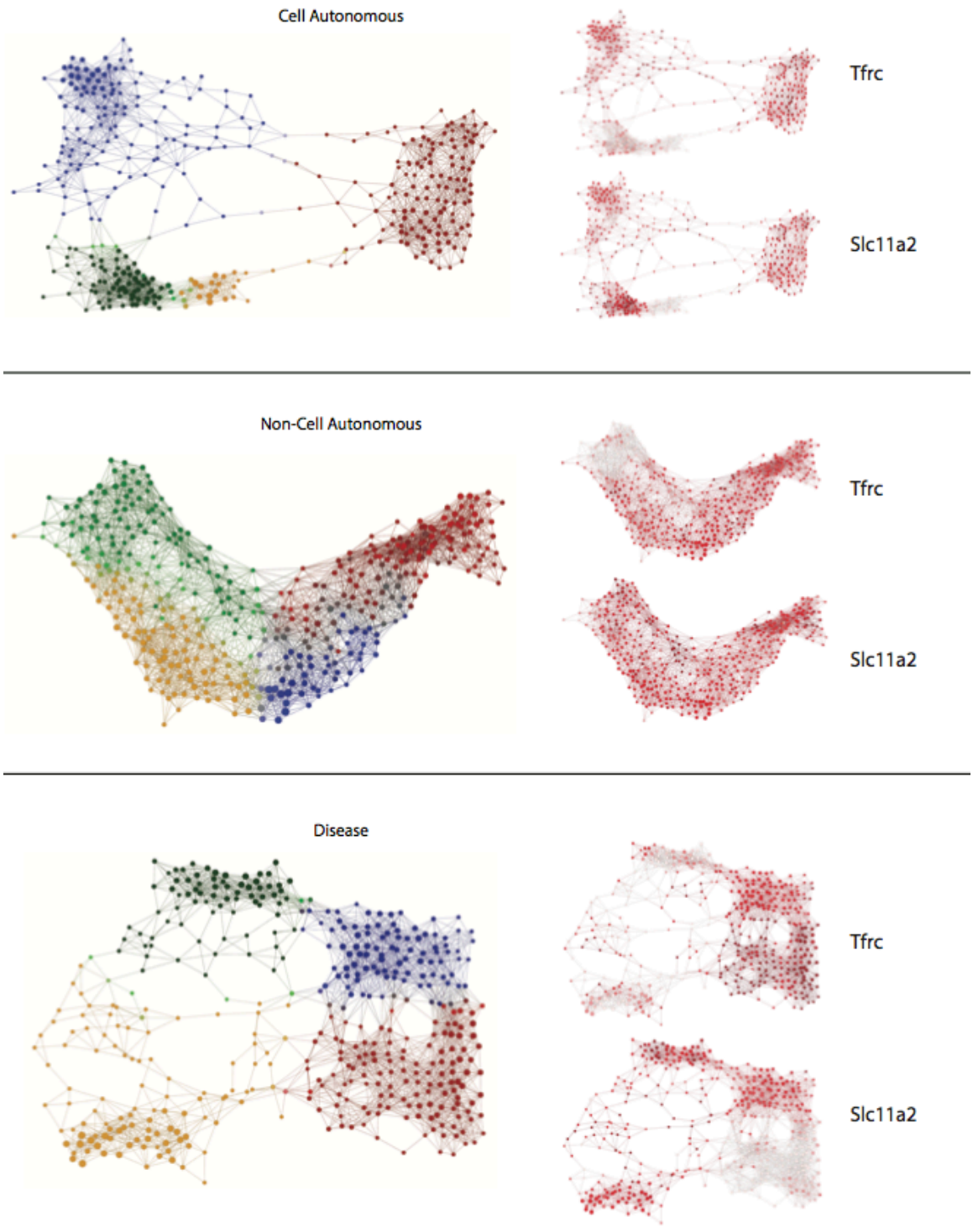


Figure 12. Changes in iron transporter expression between timepoints and conditions

Relative expression of Tfrc and Scl11a2 across conditions. (Tfrc expression log scale 6 – 7.5; Slc11a2 expression log scale 3 – 6)

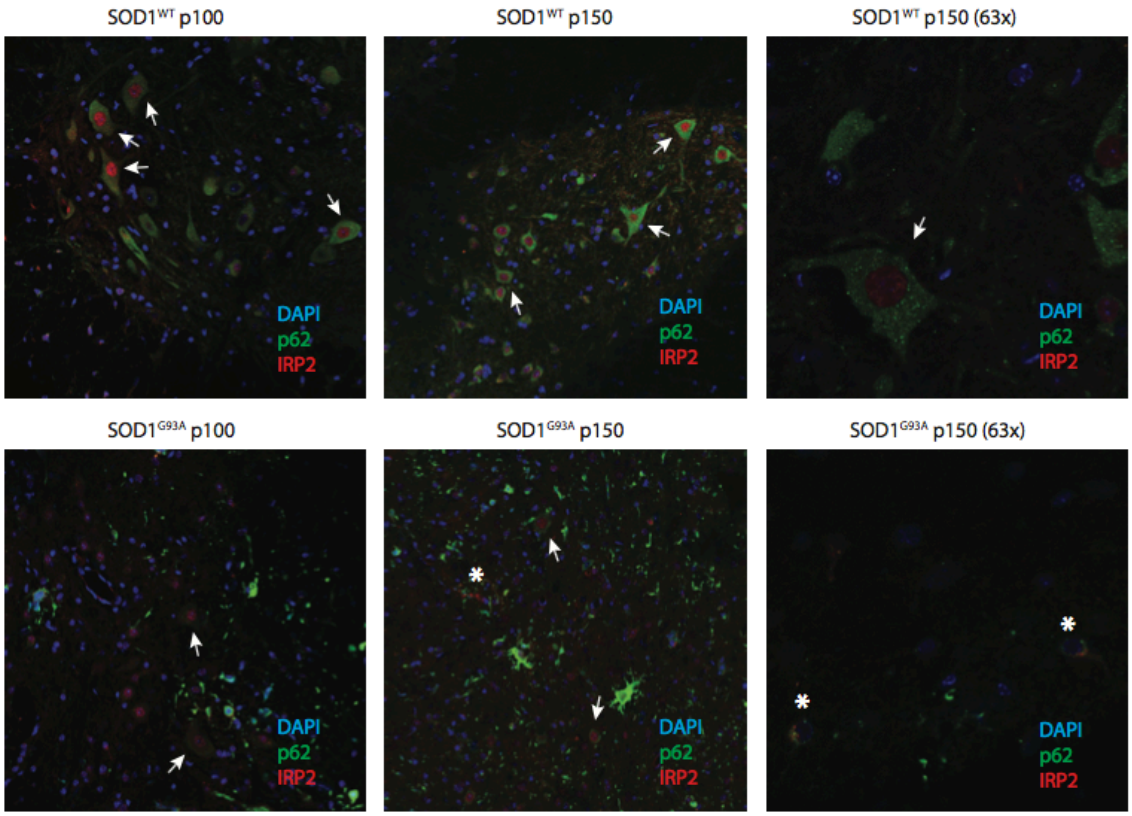


Figure 13. Loss of nuclear IRP2 in large neurons in the ventral horn of the spinal cord

IRP2 signal is largely nuclear in the large neurons of the ventral spinal cord in $SOD1^{WT}$ animals. At p100, many large nuclei in the $SOD1^{G93A}$ animal retain nuclear IRP2 signal (marked by arrowheads). At p150, most of these large nuclei are lost in the $SOD1^{G93A}$ animals and the IRP2 signal in the remaining nuclei is often localized to p62- cytoplasmic aggregates (marked by asterisk). However, low expression of p62 in cells corresponds with retained IRP2 nuclear localization (marked by arrowheads). (Figures at 20x unless otherwise noted; blue, DAPI; green, p62; red, IRP2)

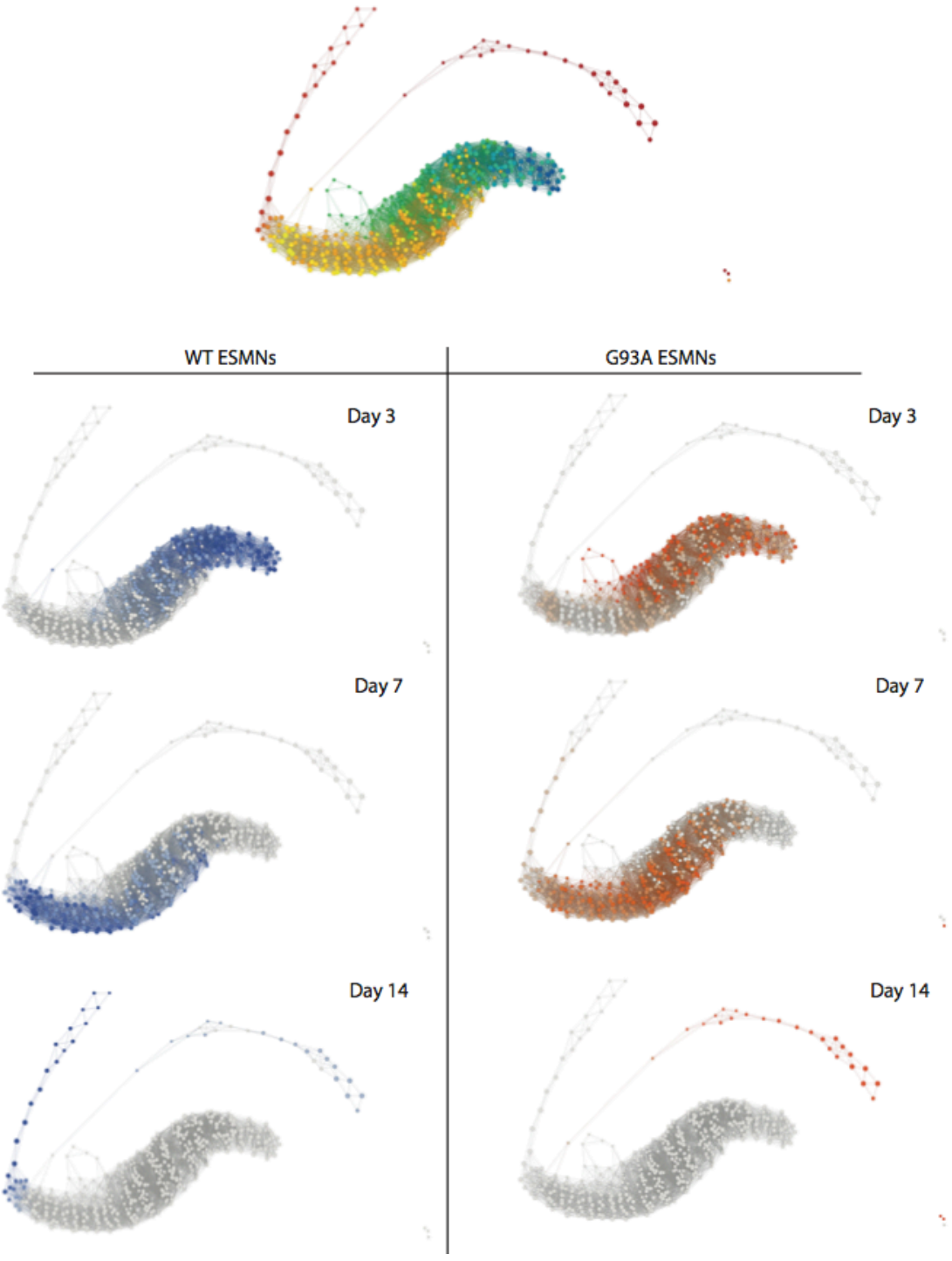


Figure 14. scTDA of SOD1^{WT} and SOD1^{G93A} ESMNs plated over WT glia.

A clear transcriptional progression is seen between SOD1^{WT} ESMNs (blue) and SOD1^{G93A} ESMNs (red) sampled at different timepoints in the topological representation.

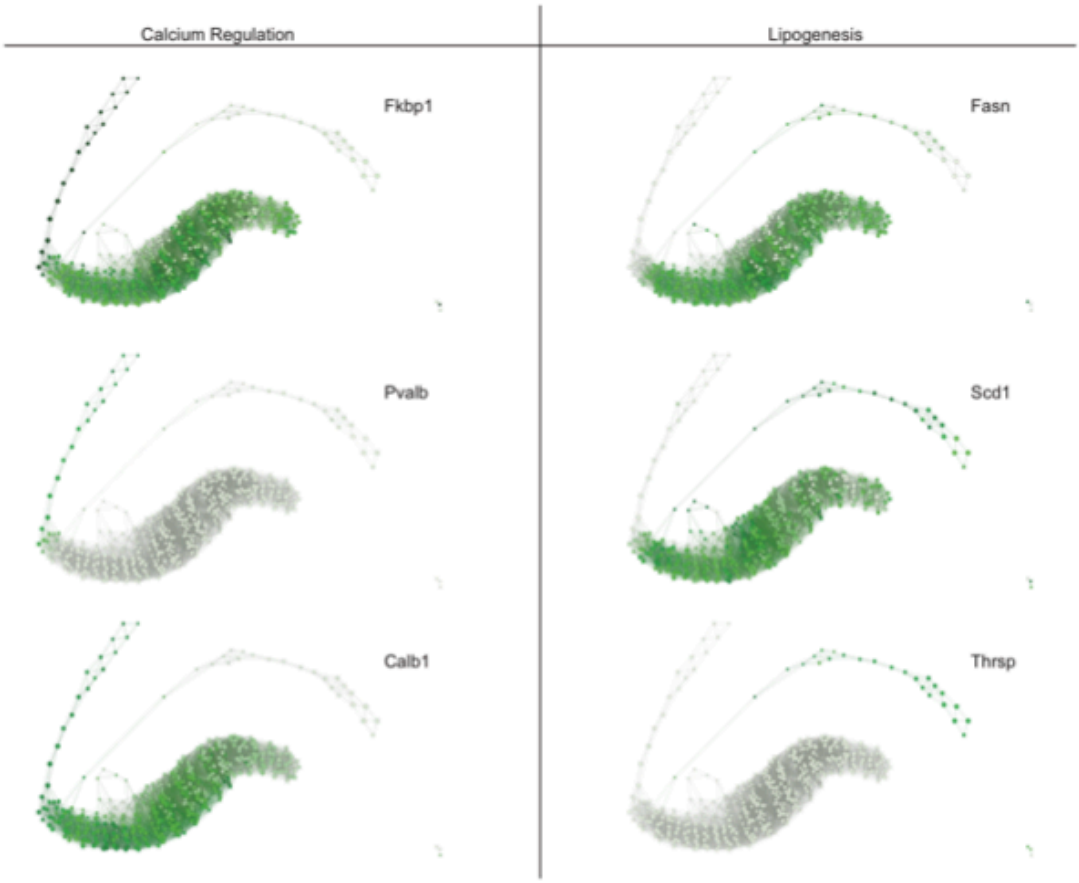


Figure 15. Two GO pathways that are alternately enriched in the two branches of day 14 ESMNs in the topological representation.

The expression of 3 genes involved in calcium regulation (left) and 3 genes involved in lipogenesis (right) are shown to be differentially enriched in the day 14 populations of predominantly $SOD1^{WT}$ (left) and $SOD1^{G93A}$ (right) ESMNs.

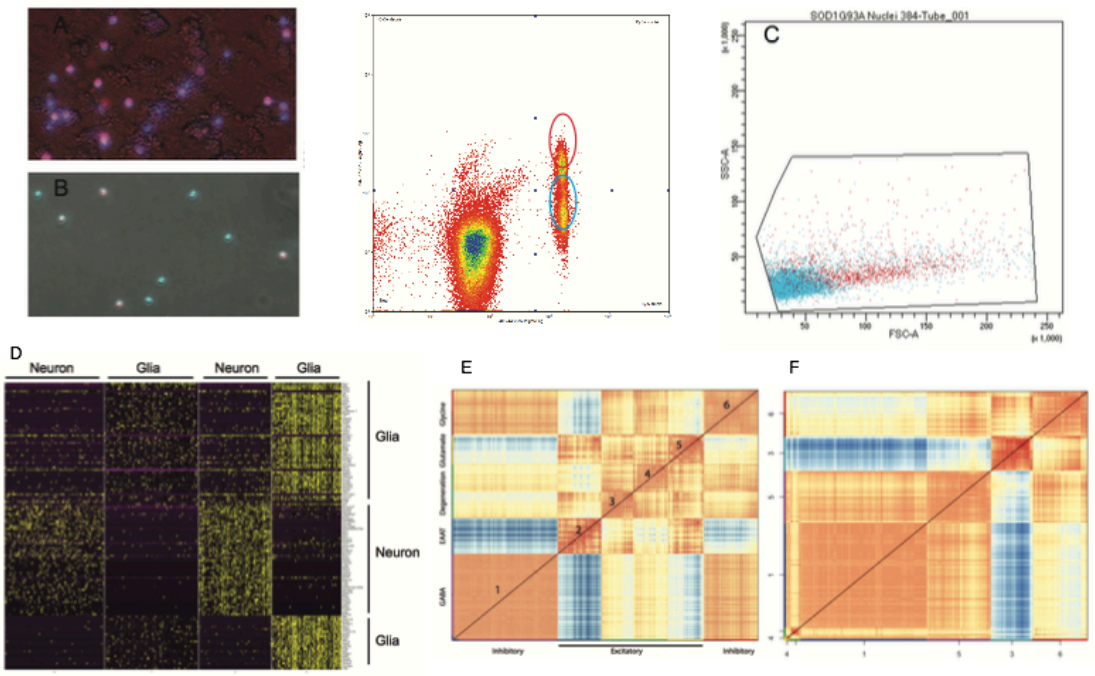


Figure 16. Pilot experiment of nuclei sequencing from the p150 SOD1^{G93A} lumbar spinal cord.

Homogenization **(A)** and purification through centrifugation **(B)** results in a clean nuclei isolation. Nuclei are labeled with DAPI. Cy3-NeuN (red) additionally labels neuronal nuclei. In the FACS sort, Cy3+/DAPI+ neuronal nuclei can be discerned from Cy3-/DAPI+ glial nuclei.

C. Consistent with expectations, on a FACS plot, Cy3+ neuronal nuclei are larger than Cy3- glial nuclei.

D. Clustering of glial and neuronal markers. 4 populations appear in the heatmap, two for neuronal nuclei and two for glial nuclei. The left neuronal and glial populations are under-sequenced compared to the populations on the right.

E. Neuronal nuclei identified by the clustering in (D) can be further defined as excitatory and inhibitory through RaceID. Clusters are identified through enriched expression of neurotransmitter transporters, and genes involved in GO pathways for DNA repair and proteosome.

F. Glial nuclei identified by the clustering in (D) can be further identified based on expression of selected marker genes as microglia (cluster 1; Selplg), oligodendrocytes (cluster 2; Plp1),

endothelial cells (cluster 3; Nrp1), astrocytes (cluster 5; Trim9), and cells undergoing stress response (cluster 6; GO pathway activation).

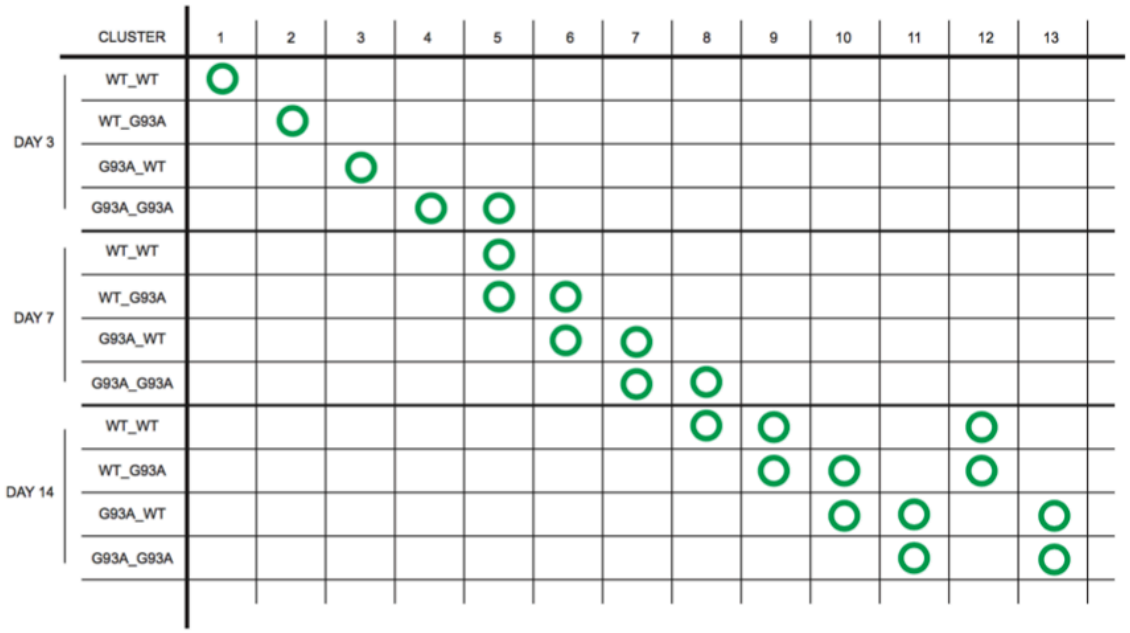


Table 1. Sampled ESMNs that populate the clusters.

ESMNs that have more than 10 cells in the cluster are marked as being present in the cluster.

References

1. Alaynick, W.A., Jessell, T.M. & Pfaff, S.L. SnapShot: spinal cord development. *Cell* **146**, 178-178 e171 (2011).

2. Jessell, T.M. Neuronal specification in the spinal cord: inductive signals and transcriptional codes. *Nature reviews. Genetics* **1**, 20-29 (2000).

3. Ericson, J. et al. Pax6 controls progenitor cell identity and neuronal fate in response to graded Shh signaling. *Cell* **90**, 169-180 (1997).

4. Lee, S.K., Lee, B., Ruiz, E.C. & Pfaff, S.L. Olig2 and Ngn2 function in opposition to modulate gene expression in motor neuron progenitor cells. *Genes Dev* **19**, 282-294 (2005).

5. Lee, S.K. & Pfaff, S.L. Synchronization of neurogenesis and motor neuron specification by direct coupling of bHLH and homeodomain transcription factors. *Neuron* **38**, 731-745 (2003).

6. Stifani, N. Motor neurons and the generation of spinal motor neuron diversity. *Frontiers in cellular neuroscience* **8**, 293 (2014).

7. Zarei, S. et al. A comprehensive review of amyotrophic lateral sclerosis. *Surgical neurology international* **6**, 171 (2015).

8. Nguyen, H.P., Van Broeckhoven, C. & van der Zee, J. ALS Genes in the Genomic Era and their Implications for FTD. *Trends in genetics : TIG* **34**, 404-423 (2018).

9. Conradi, S. & Ronnevi, L.O. Selective vulnerability of alpha motor neurons in ALS: relation to autoantibodies toward acetylcholinesterase (AChE) in ALS patients. *Brain research bulletin* **30**, 369-371 (1993).

10. Philips, T. & Rothstein, J.D. Glial cells in amyotrophic lateral sclerosis. *Experimental neurology* **262 Pt B**, 111-120 (2014).

11. Gurney, M.E. et al. Motor neuron degeneration in mice that express a human Cu,Zn superoxide dismutase mutation. *Science* **264**, 1772-1775 (1994).

12. Yamanaka, K. et al. Mutant SOD1 in cell types other than motor neurons and oligodendrocytes accelerates onset of disease in ALS mice. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 7594-7599 (2008).

13. Lino, M.M., Schneider, C. & Caroni, P. Accumulation of SOD1 mutants in postnatal motoneurons does not cause motoneuron pathology or

motoneuron disease. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **22**, 4825-4832 (2002).

14. Pramatarova, A., Laganiere, J., Roussel, J., Brisebois, K. & Rouleau, G.A. Neuron-specific expression of mutant superoxide dismutase 1 in transgenic mice does not lead to motor impairment. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **21**, 3369-3374 (2001).

15. Jaarsma, D., Teuling, E., Haasdijk, E.D., De Zeeuw, C.I. & Hoogenraad, C.C. Neuron-specific expression of mutant superoxide dismutase is sufficient to induce amyotrophic lateral sclerosis in transgenic mice. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **28**, 2075-2088 (2008).

16. Clement, A.M. et al. Wild-type nonneuronal cells extend survival of SOD1 mutant motor neurons in ALS mice. *Science* **302**, 113-117 (2003).

17. Papadeas, S.T., Kraig, S.E., O'Banion, C., Lepore, A.C. & Maragakis, N.J. Astrocytes carrying the superoxide dismutase 1 (SOD1G93A) mutation induce wild-type motor neuron degeneration in vivo. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 17803-17808 (2011).

18. Bruijn, L.I. et al. ALS-linked SOD1 mutant G85R mediates damage to astrocytes and promotes rapidly progressive disease with SOD1-containing inclusions. *Neuron* **18**, 327-338 (1997).

19. Gong, Y.H., Parsadanian, A.S., Andreeva, A., Snider, W.D. & Elliott, J.L. Restricted expression of G86R Cu/Zn superoxide dismutase in astrocytes results in astrocytosis but does not cause motoneuron degeneration. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **20**, 660-665 (2000).

20. Yamanaka, K. et al. Astrocytes as determinants of disease progression in inherited amyotrophic lateral sclerosis. *Nature neuroscience* **11**, 251-253 (2008).

21. Sargsyan, S.A., Blackburn, D.J., Barber, S.C., Monk, P.N. & Shaw, P.J. Mutant SOD1 G93A microglia have an inflammatory phenotype and elevated production of MCP-1. *Neuroreport* **20**, 1450-1455 (2009).

22. Chiu, I.M. et al. A neurodegeneration-specific gene-expression signature of acutely isolated microglia from an amyotrophic lateral sclerosis mouse model. *Cell Rep* **4**, 385-401 (2013).

23. Boillee, S. et al. Onset and progression in inherited ALS determined by motor neurons and microglia. *Science* **312**, 1389-1392 (2006).

24. Philips, T. et al. Oligodendrocyte dysfunction in the pathogenesis of amyotrophic lateral sclerosis. *Brain : a journal of neurology* **136**, 471-482 (2013).

25. Kang, S.H. et al. Degeneration and impaired regeneration of gray matter oligodendrocytes in amyotrophic lateral sclerosis. *Nature neuroscience* **16**, 571-579 (2013).

26. Zhang, Y. et al. Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals Transcriptional and Functional Differences with Mouse. *Neuron* **89**, 37-53 (2016).

27. Foo, L.C. et al. Development of a method for the purification and culture of rodent astrocytes. *Neuron* **71**, 799-811 (2011).

28. Bronstein, R., Torres, L., Nissen, J.C. & Tsirka, S.E. Culturing microglia from the neonatal and adult central nervous system. *Journal of visualized experiments : JoVE*, 50647 (2013).

29. Mizee, M.R. et al. Isolation of primary microglia from the human post-mortem brain: effects of ante- and post-mortem variables. *Acta neuropathologica communications* **5**, 16 (2017).

30. Shi, J., Marinovich, A. & Barres, B.A. Purification and characterization of adult oligodendrocyte precursor cells from the rat optic nerve. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **18**, 4627-4636 (1998).

31. Raoul, C. et al. Motoneuron death triggered by a specific pathway downstream of Fas. potentiation by ALS-linked SOD1 mutations. *Neuron* **35**, 1067-1083 (2002).

32. Aebischer, J. et al. IFNgamma triggers a LIGHT-dependent selective death of motoneurons contributing to the non-cell-autonomous effects of mutant SOD1. *Cell death and differentiation* **18**, 754-768 (2011).

33. Beaudet, M.J. et al. High yield extraction of pure spinal motor neurons, astrocytes and microglia from single embryo and adult mouse spinal cord. *Scientific reports* **5**, 16763 (2015).

34. Wichterle, H., Lieberam, I., Porter, J.A. & Jessell, T.M. Directed differentiation of embryonic stem cells into motor neurons. *Cell* **110**, 385-397 (2002).

35. Bryja, V., Bonilla, S. & Arenas, E. Derivation of mouse embryonic stem cells. *Nature protocols* **1**, 2082-2087 (2006).

36. Di Giorgio, F.P., Carrasco, M.A., Siao, M.C., Maniatis, T. & Eggan, K. Non-cell autonomous effect of glia on motor neurons in an embryonic stem cell-based ALS model. *Nature neuroscience* **10**, 608-614 (2007).

37. Wachter, N., Storch, A. & Hermann, A. Human TDP-43 and FUS selectively affect motor neuron maturation and survival in a murine cell model of ALS by non-cell-autonomous mechanisms. *Amyotrophic lateral sclerosis & frontotemporal degeneration* **16**, 431-441 (2015).

38. Biscarini, S. et al. Characterization of the lncRNA transcriptome in mESC-derived motor neurons: Implications for FUS-ALS. *Stem cell research* **27**, 172-179 (2018).

39. Shimojo, D. et al. Rapid, efficient, and simple motor neuron differentiation from human pluripotent stem cells. *Molecular brain* **8**, 79 (2015).

40. Sances, S. et al. Modeling ALS with motor neurons derived from human induced pluripotent stem cells. *Nature neuroscience* **19**, 542-553 (2016).

41. Amoroso, M.W. et al. Accelerated high-yield generation of limb-innervating motor neurons from human stem cells. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **33**, 574-586 (2013).

42. Karumbayaram, S. et al. Directed differentiation of human-induced pluripotent stem cells generates active motor neurons. *Stem cells* **27**, 806-811 (2009).

43. Liu, M.L., Zang, T. & Zhang, C.L. Direct Lineage Reprogramming Reveals Disease-Specific Phenotypes of Motor Neurons from Human ALS Patients. *Cell Rep* **14**, 115-128 (2016).

44. Kiskinis, E. et al. Pathways disrupted in human ALS motor neurons identified through genetic correction of mutant SOD1. *Cell stem cell* **14**, 781-795 (2014).

45. Devlin, A.C. et al. Human iPSC-derived motoneurons harbouring TARDBP or C9ORF72 ALS mutations are dysfunctional despite maintaining viability. *Nature communications* **6**, 5999 (2015).

46. Sareen, D. et al. Targeting RNA foci in iPSC-derived motor neurons from ALS patients with a C9ORF72 repeat expansion. *Science translational medicine* **5**, 208ra149 (2013).

47. Wainger, B.J. et al. Intrinsic membrane hyperexcitability of amyotrophic lateral sclerosis patient-derived motor neurons. *Cell Rep* **7**, 1-11 (2014).

48. Zhang, K. et al. The C9orf72 repeat expansion disrupts nucleocytoplasmic transport. *Nature* **525**, 56-61 (2015).

49. Di Giorgio, F.P., Boulting, G.L., Bobrowicz, S. & Eggan, K.C. Human embryonic stem cell-derived motor neurons are sensitive to the toxic effect of glial cells carrying an ALS-causing mutation. *Cell stem cell* **3**, 637-648 (2008).

50. Marchetto, M.C. et al. Non-cell-autonomous effect of human SOD1 G37R astrocytes on motor neurons derived from human embryonic stem cells. *Cell stem cell* **3**, 649-657 (2008).

51. Nagai, M. et al. Astrocytes expressing ALS-linked mutated SOD1 release factors selectively toxic to motor neurons. *Nature neuroscience* **10**, 615-622 (2007).

52. Varcianna, A. et al. Micro-RNAs secreted through astrocyte-derived extracellular vesicles cause neuronal network degeneration in C9orf72 ALS. *EBioMedicine* **40**, 626-635 (2019).

53. Re, D.B. et al. Necroptosis drives motor neuron death in models of both sporadic and familial ALS. *Neuron* **81**, 1001-1008 (2014).

54. Mertens, J. et al. Directly Reprogrammed Human Neurons Retain Aging-Associated Transcriptomic Signatures and Reveal Age-Related Nucleocytoplasmic Defects. *Cell stem cell* **17**, 705-718 (2015).

55. Marchetto, M.C. et al. Transcriptional signature and memory retention of human-induced pluripotent stem cells. *PloS one* **4**, e7076 (2009).

56. Davis-Dusenbery, B.N., Williams, L.A., Klim, J.R. & Eggan, K. How to make spinal motor neurons. *Development* **141**, 491-501 (2014).

57. Peljto, M., Dasen, J.S., Mazzoni, E.O., Jessell, T.M. & Wichterle, H. Functional diversity of ESC-derived motor neuron subtypes revealed through intraspinal transplantation. *Cell stem cell* **7**, 355-366 (2010).

58. Bain, G., Kitchens, D., Yao, M., Huettner, J.E. & Gottlieb, D.I. Embryonic stem cells express neuronal properties in vitro. *Developmental biology* **168**, 342-357 (1995).

59. Mazzoni, E.O. et al. Saltatory remodeling of Hox chromatin in response to rostrocaudal patterning signals. *Nature neuroscience* **16**, 1191-1198 (2013).

60. Patani, R. et al. Activin/Nodal inhibition alone accelerates highly efficient neural conversion from human embryonic stem cells and imposes a caudal positional identity. *PloS one* **4**, e7327 (2009).

61. Hjelm, B.E. et al. In vitro-differentiated neural cell cultures progress towards donor-identical brain tissue. *Human molecular genetics* **22**, 3534-3546 (2013).

62. Ho, R. et al. ALS disrupts spinal motor neuron maturation and aging pathways within gene co-expression networks. *Nature neuroscience* **19**, 1256-1267 (2016).

63. Singh Roy, N. et al. Enhancer-specified GFP-based FACS purification of human spinal motor neurons from embryonic stem cells. *Experimental neurology* **196**, 224-234 (2005).

64. Takazawa, T. et al. Maturation of spinal motor neurons derived from human embryonic stem cells. *PloS one* **7**, e40154 (2012).

65. Phatnani, H.P. et al. Intricate interplay between astrocytes and motor neurons in ALS. *Proceedings of the National Academy of Sciences of the United States of America* **110**, E756-765 (2013).

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

67. Nancy Mah et al. Evaluating Cell Identity from Transcription Profiles. *biorxiv* (2018).

68. Novick, A. & Weiner, M. Enzyme Induction as an All-or-None Phenomenon. *Proceedings of the National Academy of Sciences of the United States of America* **43**, 553-566 (1957).

69. Ko, M.S., Nakauchi, H. & Takahashi, N. The dose dependence of glucocorticoid-inducible gene expression results from changes in the number of transcriptionally active templates. *The EMBO journal* **9**, 2835-2842 (1990).

70. Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y. & Tyagi, S. Stochastic mRNA synthesis in mammalian cells. *PLoS biology* **4**, e309 (2006).

71. Thattai, M. & van Oudenaarden, A. Intrinsic noise in gene regulatory networks. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 8614-8619 (2001).

72. Balazsi, G., van Oudenaarden, A. & Collins, J.J. Cellular decision making and biological noise: from microbes to mammals. *Cell* **144**, 910-925 (2011).

73. Raj, A. & van Oudenaarden, A. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* **135**, 216-226 (2008).

74. Rizvi, A.H. et al. Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development. *Nature biotechnology* **35**, 551-560 (2017).

75. Method of the year 2013. *Nature methods* **11**, 1 (2014).

76. "Breakthrough of the year 2018". Science | AAAS. 20 December 2018.

77. Grun, D. et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525**, 251-255 (2015).

78. Shaffer, S.M. et al. Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature* **546**, 431-435 (2017).

79. Kester, L. & van Oudenaarden, A. Single-Cell Transcriptomics Meets Lineage Tracing. *Cell stem cell* **23**, 166-179 (2018).

80. Gall, J.G. & Pardue, M.L. Formation and detection of RNA-DNA hybrid molecules in cytological preparations. *Proceedings of the National Academy of Sciences of the United States of America* **63**, 378-383 (1969).

81. Zawatzky, R., De Maeyer, E. & De Maeyer-Guignard, J. Identification of individual interferon-producing cells by in situ hybridization. *Proceedings of the National Academy of Sciences of the United States of America* **82**, 1136-1140 (1985).

82. Lein, E.S. et al. Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**, 168-176 (2007).

83. van der Ploeg, M. Cytochemical nucleic acid research during the twentieth century. *European journal of histochemistry : EJH* **44**, 7-42 (2000).

84. Efstratiadis, A., Maniatis, T., Kafatos, F.C., Jeffrey, A. & Vournakis, J.N. Full length and discrete partial reverse transcripts of globin and chorion mRNAs. *Cell* **4**, 367-378 (1975).

85. Efstratiadis, A., Kafatos, F.C., Maxam, A.M. & Maniatis, T. Enzymatic in vitro synthesis of globin genes. *Cell* **7**, 279-288 (1976).

86. Eberwine, J. et al. Analysis of gene expression in single live neurons. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 3010-3014 (1992).

87. Lambolez, B., Audinat, E., Bochet, P., Crepel, F. & Rossier, J. AMPA receptor subunits expressed by single Purkinje cells. *Neuron* **9**, 247-258 (1992).

88. Mackler, S.A., Brooks, B.P. & Eberwine, J.H. Stimulus-induced coordinate changes in mRNA abundance in single postsynaptic hippocampal CA1 neurons. *Neuron* **9**, 539-548 (1992).

89. Van Gelder, R.N. et al. Amplified RNA synthesized from limited quantities of heterogeneous cDNA. *Proceedings of the National Academy of Sciences of the United States of America* **87**, 1663-1667 (1990).

90. Brady, G., Barbara, M. & Iscove, N. N. Representative in vitro cDNA amplification from individual hemopoietic cells and colonies. *Methods Mol. Cell Biol* **2**, 17-25 (1990).

91. Mullis, K. et al. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harbor symposia on quantitative biology* **51 Pt 1**, 263-273 (1986).

92. Dulac, C. & Axel, R. A novel family of genes encoding putative pheromone receptors in mammals. *Cell* **83**, 195-206 (1995).

93. Peixoto, A., Monteiro, M., Rocha, B. & Veiga-Fernandes, H. Quantification of multiple gene expression in individual cells. *Genome research* **14**, 1938-1947 (2004).

94. Emmert-Buck, M.R. et al. Laser capture microdissection. *Science* **274**, 998-1001 (1996).

95. Luo, L. et al. Gene expression profiles of laser-captured adjacent neuronal subtypes. *Nature medicine* **5**, 117-122 (1999).

96. Kamentsky, L.A. & Melamed, M.R. Spectrophotometric cell sorter. *Science* **156**, 1364-1365 (1967).

97. Picot, J., Guerin, C.L., Le Van Kim, C. & Boulanger, C.M. Flow cytometry: retrospective, fundamentals and recent instrumentation. *Cytotechnology* **64**, 109-130 (2012).

98. Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467-470 (1995).

99. Smith, L.M. et al. Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, 674-679 (1986).

100. Heather, J.M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107**, 1-8 (2016).

101. Lander, E.S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).

102. Mouse Genome Sequencing, C. et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-562 (2002).

103. Consortium, E.P. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636-640 (2004).

104. Kamme, F. et al. Single-cell microarray analysis in hippocampus CA1: demonstration and validation of cellular heterogeneity. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **23**, 3607-3615 (2003).

105. Tang, F. et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature methods* **6**, 377-382 (2009).

106. Goodwin, S., McPherson, J.D. & McCombie, W.R. Coming of age: ten years of next-generation sequencing technologies. *Nature reviews. Genetics* **17**, 333-351 (2016).

107. Costello, M. et al. Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC genomics* **19**, 332 (2018).

108. MacConaill, L.E. et al. Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. *BMC genomics* **19**, 30 (2018).

109. Head, S.R. et al. Library construction for next-generation sequencing: overviews and challenges. *BioTechniques* **56**, 61-64, 66, 68, passim (2014).

110. Parekh, S., Ziegenhain, C., Vieth, B., Enard, W. & Hellmann, I. The impact of amplification on differential expression analyses by RNA-seq. *Scientific reports* **6**, 25533 (2016).

111. Krueger, F., Andrews, S.R. & Osborne, C.S. Large scale loss of data in low-diversity illumina sequencing libraries can be recovered by deferred cluster calling. *PloS one* **6**, e16607 (2011).

112. Mitra, A., Skrzypczak, M., Ginalski, K. & Rowicka, M. Strategies for achieving high sequencing accuracy for low diversity samples and avoiding sample bleeding using illumina platform. *PloS one* **10**, e0120520 (2015).

113. Heimberg, G., Bhatnagar, R., El-Samad, H. & Thomson, M. Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing. *Cell systems* **2**, 239-250 (2016).

114. Ziegenhain, C. et al. Comparative Analysis of Single-Cell RNA Sequencing Methods. *Molecular cell* **65**, 631-643 e634 (2017).

115. Gallego Romero, I., Pai, A.A., Tung, J. & Gilad, Y. RNA-seq: impact of RNA degradation on transcript quantification. *BMC biology* **12**, 42 (2014).

116. Habib, N. et al. Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. *Science* **353**, 925-928 (2016).

117. Lake, B.B. et al. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science* **352**, 1586-1590 (2016).

118. Grindberg, R.V. et al. RNA-sequencing from single nuclei. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 19802-19807 (2013).

119. Habib, N. et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nature methods* **14**, 955-958 (2017).

120. Lacar, B. et al. Nuclear RNA-seq of single neurons reveals molecular signatures of activation. *Nature communications* **7**, 11022 (2016).

121. Lake, B.B. et al. A comparative strategy for single-nucleus and single-cell transcriptomes confirms accuracy in predicted cell-type expression from nuclear RNA. *Scientific reports* **7**, 6031 (2017).

122. La Manno, G. et al. RNA velocity of single cells. *Nature* **560**, 494-498 (2018).

123. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nature protocols* **9**, 171-181 (2014).

124. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell Reports* **2**, 666-673 (2012).

125. Hashimshony, T. et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome biology* **17**, 77 (2016).

126. Bagnoli, J.W. et al. Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. *Nature communications* **9**, 2937 (2018).

127. Rosenberg, A.B. et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176-182 (2018).

128. Macosko, E.Z. et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214 (2015).

129. Klein, A.M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187-1201 (2015).

130. Pollen, A.A. et al. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nature biotechnology* **32**, 1053-1058 (2014).

131. Zheng, G.X. et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications* **8**, 14049 (2017).

132. Lodish, H. Molecular cell biology. (W. H. Freeman and Co., New York; 2016).

133. Livesey, F.J. Strategies for microarray analysis of limiting amounts of RNA. *Briefings in functional genomics & proteomics* **2**, 31-36 (2003).

134. Islam, S. et al. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature methods* **11**, 163-166 (2014).

135. Matz, M. et al. Amplification of cDNA ends based on template-switching effect and step-out PCR. *Nucleic acids research* **27**, 1558-1560 (1999).

136. Zhu, Y.Y., Machleder, E.M., Chenchik, A., Li, R. & Siebert, P.D. Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *BioTechniques* **30**, 892-897 (2001).

137. Luo, G.X. & Taylor, J. Template switching by reverse transcriptase during DNA synthesis. *Journal of virology* **64**, 4321-4328 (1990).

138. Chen, D. & Patton, J.T. Reverse transcriptase adds nontemplated nucleotides to cDNAs during 5'-RACE and primer extension. *BioTechniques* **30**, 574-580, 582 (2001).

139. Schmidt, W.M. & Mueller, M.W. CapSelect: a highly sensitive method for 5' CAP-dependent enrichment of full-length cDNA in PCR-mediated analysis of mRNAs. *Nucleic acids research* **27**, e31 (1999).

140. Zhou, Y. & Martin, C.T. Observed instability of T7 RNA polymerase elongation complexes can be dominated by collision-induced "bumping". *The Journal of biological chemistry* **281**, 24441-24448 (2006).

141. Bustin, S. et al. Variability of the reverse transcription step: practical implications. *Clinical chemistry* **61**, 202-212 (2015).

142. Picelli, S. et al. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome research* **24**, 2033-2040 (2014).

143. Suzuki, M.T. & Giovannoni, S.J. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Applied and environmental microbiology* **62**, 625-630 (1996).

144. Polz, M.F. & Cavanaugh, C.M. Bias in template-to-product ratios in multitemplate PCR. *Applied and environmental microbiology* **64**, 3724-3730 (1998).

145. Eckert, K.A. & Kunkel, T.A. DNA polymerase fidelity and the polymerase chain reaction. *PCR methods and applications* **1**, 17-24 (1991).

146. Svensson, V. et al. Power analysis of single-cell RNA-sequencing experiments. *Nature methods* **14**, 381-387 (2017).

147. Smyth, R.P. et al. Reducing chimera formation during PCR amplification to ensure accurate genotyping. *Gene* **469**, 45-51 (2010).

148. Levin, J.Z. et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nature methods* **7**, 709-715 (2010).

149. Song, Y., Liu, K.J. & Wang, T.H. Elimination of ligation dependent artifacts in T4 RNA ligase to achieve high efficiency and low bias microRNA capture. *PloS one* **9**, e94619 (2014).

150. Jayaprakash, A.D., Jabado, O., Brown, B.D. & Sachidanandam, R. Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. *Nucleic acids research* **39**, e141 (2011).

151. Gyarmati, P., Song, Y., Hallman, J. & Kaller, M. Chemical fragmentation for massively parallel sequencing library preparation. *Journal of biotechnology* **168**, 95-100 (2013).

152. Adey, A. et al. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome biology* **11**, R119 (2010).

153. Knierim, E., Lucke, B., Schwarz, J.M., Schuelke, M. & Seelow, D. Systematic comparison of three methods for fragmentation of long-range PCR products for next generation sequencing. *PloS one* **6**, e28240 (2011).

154. Kia, A. et al. Improved genome sequencing using an engineered transposase. *BMC biotechnology* **17**, 6 (2017).

155. Goh, W.W.B., Wang, W. & Wong, L. Why Batch Effects Matter in Omics Data, and How to Avoid Them. *Trends in biotechnology* **35**, 498-507 (2017).

156. Faridani, O.R. et al. Single-cell sequencing of the small-RNA transcriptome. *Nature biotechnology* **34**, 1264-1266 (2016).

157. Fang, N. & Akinci-Tolun, R. Depletion of Ribosomal RNA Sequences from Single-Cell RNA-Sequencing Library. *Current protocols in molecular biology* **115**, 7 27 21-27 27 20 (2016).

158. Zhang, Z. et al. Enhanced amplification of GC-rich DNA with two organic reagents. *BioTechniques* **47**, 775-779 (2009).

159. Zajac, P., Islam, S., Hochgerner, H., Lonnerberg, P. & Linnarsson, S. Base preferences in non-templated nucleotide incorporation by MMLV-derived reverse transcriptases. *PloS one* **8**, e85270 (2013).

160. Owczarzy, R., You, Y., Groth, C.L. & Tataurov, A.V. Stability and mismatch discrimination of locked nucleic acid-DNA duplexes. *Biochemistry* **50**, 9352-9367 (2011).

161. Minoche, A.E., Dohm, J.C. & Himmelbauer, H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome biology* **12**, R112 (2011).

162. Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome research* **27**, 491-499 (2017).

163. Tian, L. et al. scPipe: A flexible R/Bioconductor preprocessing pipeline for single-cell RNA-sequencing data. *PLoS computational biology* **14**, e1006361 (2018).

164. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10**, R25 (2009).

165. Keel, B.N. & Snelling, W.M. Comparison of Burrows-Wheeler Transform-Based Mapping Algorithms Used in High-Throughput Whole-Genome Sequencing: Application to Illumina Data for Livestock Genomes. *Frontiers in genetics* **9**, 35 (2018).

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

167. Bray, N.L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology* **34**, 525-527 (2016).

168. Cunningham, F. et al. Ensembl 2019. *Nucleic acids research* (2018).

169. Hicks, S.C., Townes, F.W., Teng, M. & Irizarry, R.A. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* **19**, 562-578 (2018).

170. Kim, T. et al. Impact of similarity metrics on single-cell RNA-seq data clustering. *Briefings in bioinformatics* (2018).

171. Lum, P.Y. et al. Extracting insights from the shape of complex data using topology. *Scientific reports* **3**, 1236 (2013).

172. Nielson, J.L. et al. Topological data analysis for discovery in preclinical spinal cord injury and traumatic brain injury. *Nature communications* **6**, 8581 (2015).

173. Guo, M. et al. Single cell RNA analysis identifies cellular heterogeneity and adaptive responses of the lung at birth. *Nature communications* **10**, 37 (2019).

174. Jolliffe, I. Principal component analysis. *Wiley Online Library* (2002).

175. Buettner, F. et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature biotechnology* **33**, 155-160 (2015).

176. Nesterov, Y., Nemirovskii, A. & Ye, Y. Interior-point polynomial algorithms in convex programming. *SIAM* **13** (1994).

177. Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research* **9**, 85 (2008).

178. Tung, P.Y. et al. Batch effects and the effective design of single-cell gene expression studies. *Scientific reports* **7**, 39921 (2017).

179. Boufea, K., Seth, S. & Batada, N.N. scID: Identification of equivalent transcriptional cell populations across single cell RNA-seq data using discriminant analysis. *biorxiv* (2019).

180. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology* **36**, 411-420 (2018).

181. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology* **32**, 381-386 (2014).

182. Qiu, X. et al. Reversed graph embedding resolves complex single-cell trajectories. *Nature methods* **14**, 979-982 (2017).

183. Haghverdi, L., Buttner, M., Wolf, F.A., Buettner, F. & Theis, F.J. Diffusion pseudotime robustly reconstructs lineage branching. *Nature methods* **13**, 845-848 (2016).

184. Welch, J.D., Hartemink, A.J. & Prins, J.F. SLICER: inferring branched, nonlinear cellular trajectories from single cell RNA-seq data. *Genome biology* **17**, 106 (2016).

185. Setty, M. et al. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nature biotechnology* **34**, 637-645 (2016).

186. Edelsbrunner, H., Letscher, D. & Zomorodian, A. Topological persistence and simplification. *Discrete Comput. Geom.* **28**, 511–533 (2002).

187. Stahl, P.L. et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78-82 (2016).

188. Strell, C. et al. Placing RNA in context and space - methods for spatially resolved transcriptomics. *The FEBS journal* (2018).

189. Behjati Ardakani, F. et al. Integrative analysis of single-cell expression data reveals distinct regulatory states in bidirectional promoters. *Epigenetics & chromatin* **11**, 66 (2018).

190. Cusanovich, D.A. et al. A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell* **174**, 1309-1324 e1318 (2018).

191. Bratt-Leal, A.M., Carpenedo, R.L. & McDevitt, T.C. Engineering the embryoid body microenvironment to direct embryonic stem cell differentiation. *Biotechnology progress* **25**, 43-51 (2009).

192. Angerer, P. et al. destiny: diffusion maps for large-scale single-cell data in R. *Bioinformatics* **32**, 1241-1243 (2016).

193. Marco, E. et al. Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proceedings of the National Academy of Sciences of the United States of America* **111**, E5643-5650 (2014).

194. Chan, J.M., Carlsson, G. & Rabadan, R. Topology of viral evolution. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 18566-18571 (2013).

195. Camara, P.G., Levine, A.J. & Rabadan, R. Inference of Ancestral Recombination Graphs through Topological Data Analysis. *PLoS computational biology* **12**, e1005071 (2016).

196. Camara, P.G., Rosenbloom, D.I., Emmett, K.J., Levine, A.J. & Rabadan, R. Topological Data Analysis Generates High-Resolution, Genome-wide Maps of Human Recombination. *Cell systems* **3**, 83-94 (2016).

197. Nicolau, M., Levine, A.J. & Carlsson, G. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 7265-7270 (2011).

198. Li, L. et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Science translational medicine* **7**, 311ra174 (2015).

199. Bendall, S.C. et al. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157**, 714-725 (2014).

200. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nature biotechnology* **33**, 495-502 (2015).

201. Singh, G., Memoli, F. & Carlsson, G.E. in SPBG 91-100. *Citeseer* (2007).

202. Finak, G. et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome biology* **16**, 278 (2015).

203. McDavid, A., Finak, G. & Gottardo, R. The contribution of cell cycle to heterogeneity in single-cell RNA-seq data. *Nature biotechnology* **34**, 591-593 (2016).

204. Balmer, J.E. & Blomhoff, R. Gene expression regulation by retinoic acid. *Journal of lipid research* **43**, 1773-1808 (2002).

205. Rhinn, M. & Dolle, P. Retinoic acid signalling during development. *Development* **139**, 843-858 (2012).

206. Gaunt, S.J. & Strachan, L. Temporal colinearity in expression of anterior Hox genes in developing chick embryos. *Developmental dynamics : an official publication of the American Association of Anatomists* **207**, 270-280 (1996).

207. Zhang, X., Weissman, S.M. & Newburger, P.E. Long intergenic non-coding RNA HOTAIRM1 regulates cell cycle progression during myeloid maturation in NB4 human promyelocytic leukemia cells. *RNA biology* **11**, 777-787 (2014).

208. Lin, M. et al. RNA-Seq of human neurons derived from iPS cells reveals candidate long non-coding RNAs involved in neurogenesis and neuropsychiatric disorders. *PloS one* **6**, e23356 (2011).

209. Mallo, M. & Alonso, C.R. The regulation of Hox gene expression during animal development. *Development* **140**, 3951-3963 (2013).

210. Dinger, M.E. et al. Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome research* **18**, 1433-1445 (2008).

211. Sommer, L., Ma, Q. & Anderson, D.J. neurogenins, a novel family of atonal-related bHLH transcription factors, are putative mammalian neuronal determination genes that reveal progenitor cell heterogeneity in the developing CNS and PNS. *Molecular and cellular neurosciences* **8**, 221-241 (1996).

212. Darnell, R.B. RNA protein interaction in neurons. *Annual review of neuroscience* **36**, 243-270 (2013).

213. Quesnel-Vallieres, M., Irimia, M., Cordes, S.P. & Blencowe, B.J. Essential roles for the splicing regulator nSR100/SRRM4 during nervous system development. *Genes & development* **29**, 746-759 (2015).

214. Calarco, J.A. et al. Regulation of vertebrate nervous system alternative splicing and development by an SR-related protein. *Cell* **138**, 898-910 (2009).

215. Treutlein, B. et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509**, 371-375 (2014).

216. Petropoulos, S. et al. Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. *Cell* **167**, 285 (2016).

217. Telley, L. et al. Sequential transcriptional waves direct the differentiation of newborn neurons in the mouse neocortex. *Science* **351**, 1443-1446 (2016).

218. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111 (2009).

219. Anders, S., Pyl, P.T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166-169 (2015).

220. Stegle, O., Teichmann, S.A. & Marioni, J.C. Computational and analytical challenges in single-cell transcriptomics. *Nature reviews. Genetics* **16**, 133-145 (2015).

221. Grun, D., Kester, L. & van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nature methods* **11**, 637-640 (2014).

222. Kharchenko, P.V., Silberstein, L. & Scadden, D.T. Bayesian approach to single-cell differential expression analysis. *Nature methods* **11**, 740-742 (2014).

223. Shalek, A.K. et al. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510**, 363-369 (2014).

224. Brennecke, P. et al. Accounting for technical noise in single-cell RNA-seq experiments. *Nature methods* **10**, 1093-1095 (2013).

225. Edelsbrunner, &, L. & Zomorodian Topological Persistence and Simplification. *Discrete & Computational Geometry*, 511–533 (2002).

226. Zomorodian, A. & & Carlsson, G. Computing persistent homology. *Discrete & Computational Geometry* **33**, 249-274 (2005).

227. Binns, D. et al. QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics* **25**, 3045-3046 (2009).

228. UniProt, C. UniProt: a hub for protein information. *Nucleic acids research* **43**, D204-212 (2015).

229. Mi, H., Muruganujan, A., Casagrande, J.T. & Thomas, P.D. Large-scale gene function analysis with the PANTHER classification system. *Nature protocols* **8**, 1551-1566 (2013).

230. Zhao, Y. et al. NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic acids research* **44**, D203-208 (2016).

231. Li, H. et al. Classifying Drosophila Olfactory Projection Neuron Subtypes by Single-Cell RNA Sequencing. *Cell* **171**, 1206-1220 e1222 (2017).

232. Chan, T.E., Stumpf, M.P.H. & Babtie, A.C. Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures. *Cell systems* **5**, 251-267 e253 (2017).

233. Van Damme, P., Robberecht, W. & Van Den Bosch, L. Modelling amyotrophic lateral sclerosis: progress and possibilities. *Disease models & mechanisms* **10**, 537-549 (2017).

234. Ilary Allodi et al. Modeling motor neuron resilience in ALS using stem cells. *biorxiv* (2019).

235. Meyer, K. et al. Direct conversion of patient fibroblasts demonstrates non-cell autonomous toxicity of astrocytes to motor neurons in familial and

sporadic ALS. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 829-832 (2014).

236. Fischer, L.R. & Glass, J.D. Axonal degeneration in motor neuron disease. *Neuro-degenerative diseases* **4**, 431-442 (2007).

237. Nijssen, J., Aguila, J., Hoogstraaten, R., Kee, N. & Hedlund, E. Axon-Seq Decodes the Motor Axon Transcriptome and Its Modulation in Response to ALS. *Stem cell reports* **11**, 1565-1578 (2018).

238. Kaplan, A. et al. Neuronal matrix metalloproteinase-9 is a determinant of selective neurodegeneration. *Neuron* **81**, 333-348 (2014).

239. Comley, L. et al. Motor neurons with differential vulnerability to degeneration show distinct protein signatures in health and ALS. *Neuroscience* **291**, 216-229 (2015).

240. Weinreb, C., Wolock, S. & Klein, A.M. SPRING: a kinetic interface for visualizing high dimensional single-cell expression data. *Bioinformatics* **34**, 1246-1248 (2018).

241. Landry, C.F. et al. Myelin basic protein gene expression in neurons: developmental and regional changes in protein targeting within neuronal nuclei, cell bodies, and processes. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **16**, 2452-2462 (1996).

242. Pribyl, T.M. et al. Expression of the myelin basic protein gene locus in neurons and oligodendrocytes in the human fetal central nervous system. *The Journal of comparative neurology* **374**, 342-353 (1996).

243. Bongarzone, E.R. et al. Two neuronal cell lines expressing the myelin basic protein gene display differences in their in vitro survival and in their response to glia. *Journal of neuroscience research* **54**, 309-319 (1998).

244. Hol, E.M. et al. Neuronal expression of GFAP in patients with Alzheimer pathology and identification of novel GFAP splice forms. *Molecular psychiatry* **8**, 786-796 (2003).

245. André Varcianna et al. Micro-RNAs secreted through astrocyte-derived extracellular vesicles cause neuronal network degeneration in C9orf72 ALS. *EBioMedicine* (2019).

246. Basso, M. et al. Mutant copper-zinc superoxide dismutase (SOD1) induces protein secretion pathway alterations and exosome release in astrocytes: implications for disease spreading and motor neuron pathology in amyotrophic lateral sclerosis. *The Journal of biological chemistry* **288**, 15699-15711 (2013).

247. Ferrara, D., Pasetto, L., Bonetto, V. & Basso, M. Role of Extracellular Vesicles in Amyotrophic Lateral Sclerosis. *Frontiers in neuroscience* **12**, 574 (2018).

248. Weber, B. & Barros, L.F. The Astrocyte: Powerhouse and Recycling Center. *Cold Spring Harbor perspectives in biology* **7** (2015).

249. Soreq, L. et al. Major Shifts in Glial Regional Identity Are a Transcriptional Hallmark of Human Brain Aging. *Cell Rep* **18**, 557-570 (2017).

250. Oberacker, T. et al. Enhanced expression of thioredoxin-interacting-protein regulates oxidative DNA damage and aging. *FEBS letters* **592**, 2297-2307 (2018).

251. Yoshida, T., Nakamura, H., Masutani, H. & Yodoi, J. The involvement of thioredoxin and thioredoxin binding protein-2 on cellular proliferation and aging process. *Annals of the New York Academy of Sciences* **1055**, 1-12 (2005).

252. Kim, G.S., Jung, J.E., Narasimhan, P., Sakata, H. & Chan, P.H. Induction of thioredoxin-interacting protein is mediated by oxidative stress, calcium, and glucose after brain injury in mice. *Neurobiology of disease* **46**, 440-449 (2012).

253. Maniatis, S. et al. Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis. *Science* **364**, 89-93 (2019).

254. Barber, S.C. & Shaw, P.J. Oxidative stress in ALS: key role in motor neuron injury and therapeutic target. *Free radical biology & medicine* **48**, 629-641 (2010).

255. Anthony, T.G. & Wek, R.C. TXNIP switches tracks toward a terminal UPR. *Cell metabolism* **16**, 135-137 (2012).

256. Prins, D. & Michalak, M. Organellar calcium buffers. *Cold Spring Harbor perspectives in biology* **3** (2011).

257. Oliet, S.H., Piet, R. & Poulain, D.A. Control of glutamate clearance and synaptic efficacy by glial coverage of neurons. *Science* **292**, 923-926 (2001).

258. Brzoska, K., Meczynska, S. & Kruszewski, M. Iron-sulfur cluster proteins: electron transfer and beyond. *Acta biochimica Polonica* **53**, 685-691 (2006).

259. Emerit, J., Beaumont, C. & Trivin, F. Iron metabolism, free radicals, and oxidative injury. *Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie* **55**, 333-339 (2001).

260. Li, K. & Reichmann, H. Role of iron in neurodegenerative diseases. *Journal of neural transmission* **123**, 389-399 (2016).

261. Bandyopadhyay, U. et al. RNA-Seq profiling of spinal cord motor neurons from a presymptomatic SOD1 ALS mouse. *PloS one* **8**, e53575 (2013).

262. Henderson, B.R., Menotti, E. & Kuhn, L.C. Iron regulatory proteins 1 and 2 bind distinct sets of RNA target sequences. *The Journal of biological chemistry* **271**, 4900-4908 (1996).

263. Ghosh, M.C., Zhang, D.L. & Rouault, T.A. Iron misregulation and neurodegenerative disease in mouse models that lack iron regulatory proteins. *Neurobiology of disease* **81**, 66-75 (2015).

264. Jeong, S.Y. et al. Iron insufficiency compromises motor neurons and their mitochondrial function in Irp2-null mice. *PloS one* **6**, e25404 (2011).

265. Magali Soumillon, Davide Cacchiarelli, Stefan Semrau, Alexander van Oudenaarden & Mikkelsen, T.S. Characterization of directed differentiation by high-throughput single-cell RNA-Seq. *biorxiv* (2014).

266. Moreau, C. et al. Could Conservative Iron Chelation Lead to Neuroprotection in Amyotrophic Lateral Sclerosis? *Antioxidants & redox signaling* **29**, 742-748 (2018).

267. Gajowiak, A., Stys, A., Starzynski, R.R., Staron, R. & Lipinski, P. Misregulation of iron homeostasis in amyotrophic lateral sclerosis. *Postepy higieny i medycyny doswiadczalnej* **70**, 709-721 (2016).

268. Jaronen, M., Goldsteins, G. & Koistinaho, J. ER stress and unfolded protein response in amyotrophic lateral sclerosis-a controversial role of protein disulphide isomerase. *Frontiers in cellular neuroscience* **8**, 402 (2014).

269. Hare, D., Ayton, S., Bush, A. & Lei, P. A delicate balance: Iron metabolism and diseases of the brain. *Frontiers in aging neuroscience* **5**, 34 (2013).

270. Gunshin, H. et al. Iron-dependent regulation of the divalent metal ion transporter. *FEBS letters* **509**, 309-316 (2001).

271. Koeller, D.M. et al. A cytosolic protein binds to structural elements within the iron regulatory region of the transferrin receptor mRNA. *Proceedings of the National Academy of Sciences of the United States of America* **86**, 3574-3578 (1989).

272. Shi, P., Gal, J., Kwinter, D.M., Liu, X. & Zhu, H. Mitochondrial dysfunction in amyotrophic lateral sclerosis. *Biochimica et biophysica acta* **1802**, 45-51 (2010).

273. Bernard-Marissal, N., Chrast, R. & Schneider, B.L. Endoplasmic reticulum and mitochondria in diseases of motor and sensory neurons: a broken relationship? *Cell death & disease* **9**, 333 (2018).

274. Niccoli, T., Partridge, L. & Isaacs, A.M. Ageing as a risk factor for ALS/FTD. *Human molecular genetics* **26**, R105-R113 (2017).

275. Nordberg, J. & Arner, E.S. Reactive oxygen species, antioxidants, and the mammalian thioredoxin system. *Free radical biology & medicine* **31**, 1287-1312 (2001).

276. Godoy, J.R. et al. Redox atlas of the mouse. Immunohistochemical detection of glutaredoxin-, peroxiredoxin-, and thioredoxin-family proteins in various tissues of the laboratory mouse. *Biochimica et biophysica acta* **1810**, 2-92 (2011).

277. Stemme, S., Hansson, H.A., Holmgren, A. & Rozell, B. Axoplasmic transport of thioredoxin and thioredoxin reductase in rat sciatic nerve. *Brain research* **359**, 140-146 (1985).

278. Lippoldt, A. et al. Localization of thioredoxin in the rat brain and functional implications. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **15**, 6747-6756 (1995).

279. Hori, K. et al. Neuroprotection by glial cells through adult T cell leukemia-derived factor/human thioredoxin (ADF/TRX). *Brain research* **652**, 304-310 (1994).

280. Malaspina, A., Kaushik, N. & de Belleroche, J. Differential expression of 14 genes in amyotrophic lateral sclerosis spinal cord detected using gridded cDNA arrays. *Journal of neurochemistry* **77**, 132-145 (2001).

281. Hwang, J. et al. The structural basis for the negative regulation of thioredoxin by thioredoxin-interacting protein. *Nature communications* **5**, 2958 (2014).

282. Oslowski, C.M. et al. Thioredoxin-interacting protein mediates ER stress-induced beta cell death through initiation of the inflammasome. *Cell metabolism* **16**, 265-273 (2012).

283. Lerner, A.G. et al. IRE1alpha induces thioredoxin-interacting protein to activate the NLRP3 inflammasome and promote programmed cell death under irremediable ER stress. *Cell metabolism* **16**, 250-264 (2012).

284. Kanouchi, T., Ohkubo, T. & Yokota, T. Can regional spreading of amyotrophic lateral sclerosis motor symptoms be explained by prion-like propagation? *Journal of neurology, neurosurgery, and psychiatry* **83**, 739-745 (2012).

285. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nature methods* **14**, 865-868 (2017).

286. Cao, J. et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380-1385 (2018).

287. Crosetto, N., Bienko, M. & van Oudenaarden, A. Spatially resolved transcriptomics and beyond. *Nature reviews. Genetics* **16**, 57-66 (2015).

288. Turner, M.R. et al. Pattern of spread and prognosis in lower limb-onset ALS. *Amyotrophic lateral sclerosis : official publication of the World Federation of Neurology Research Group on Motor Neuron Diseases* **11**, 369-373 (2010).

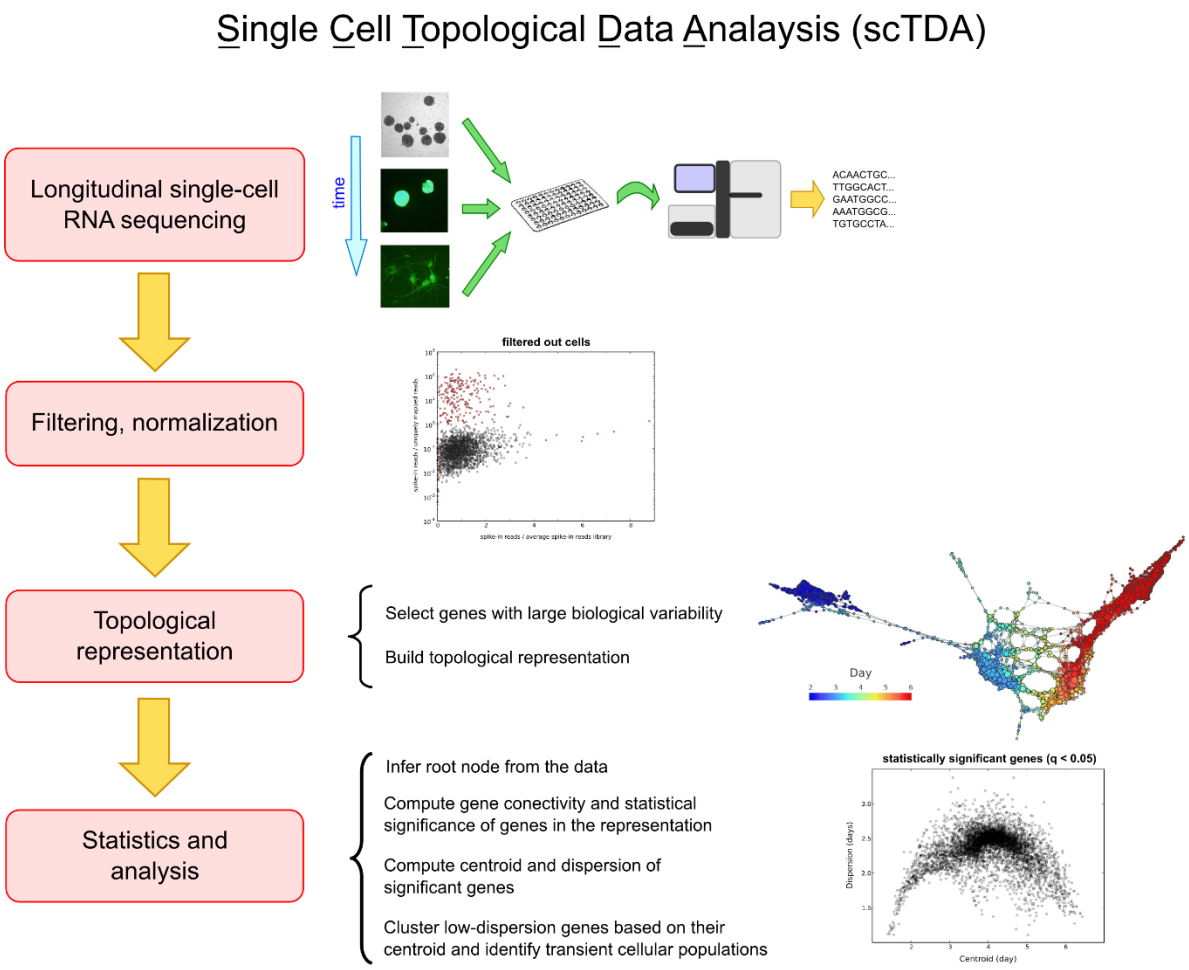
289. Ravits, J.M. & La Spada, A.R. ALS motor phenotype heterogeneity, focality, and spread: deconstructing motor neuron degeneration. *Neurology* **73**, 805-811 (2009).

Supplementary Note 1

Single Cell Library Generation

In one biological replicate, we sorted a small sample size from a differentiation, sequencing 80 cells per differentiation time-point utilizing standard CEL-Seq primers with anchor bases at the 3' end of reverse transcription primers, pooling 40 cells at a time prior to *in vitro* transcription (IVT). To assess library saturation and capture efficiency, two single cell libraries from each differentiation time point (consisting of 40 cells each) from the pilot experiment were paired end sequenced (2x125 bps) on an Illumina HiSeq 2500, operating in high output mode, sequencing with Illumina v4 chemistry. To increase capture efficiency, enhanced *in vitro* transcription based amplification, and leveraging the library saturation curves from our pilot experiment, we utilized 96 barcoded CEL-Seq RT primers (**Supplementary Table 4**), forgoing the usage of anchor bases at the 3' terminus. We then conducted a differentiation on a second biological replicate, sampling 384 cells per time-point (inclusive of 96 FACS purified mid-level GFP expressing, and 288 high GFP expressing cells), collected into 96 well plates and implemented CEL-Seq, now pooling 96 cells per IVT reaction. Following IVT, aRNA was fragmented using magnesium (NEBNext Magnesium RNA Fragmentation Module) for 90 seconds and column purified (Zymo Research RNA Clean & Concentrator-5). Purified aRNA was then subjected to treatment with Antarctic Phosphatase and T4 polynucleotide kinase. Ligation of Illumina RA3 adapters was conducted using truncated T4 RNA Ligase 2 for 1 hour at 28 C. Following adapter ligation, adapter ligated aRNA was reverse transcribed using Illumina RTP at 50 C for 1 hour and placed on ice. To avoid amplification based batch effects, the resultant cDNA was PCR amplified with Illumina RPIX primers to no more than 15 cycles. The sequencing libraries were then twice purified using AmpureXP beads, held at a ratio of 1:0.65, yielding size selected libraries with an insert size of ~250 bps. The single cell libraries were then multiplexed to a total representation of 384 cells per lane at equimolar concentrations and mixed with 50% exome libraries generated using an Illumina TruSeq Exome Kit.

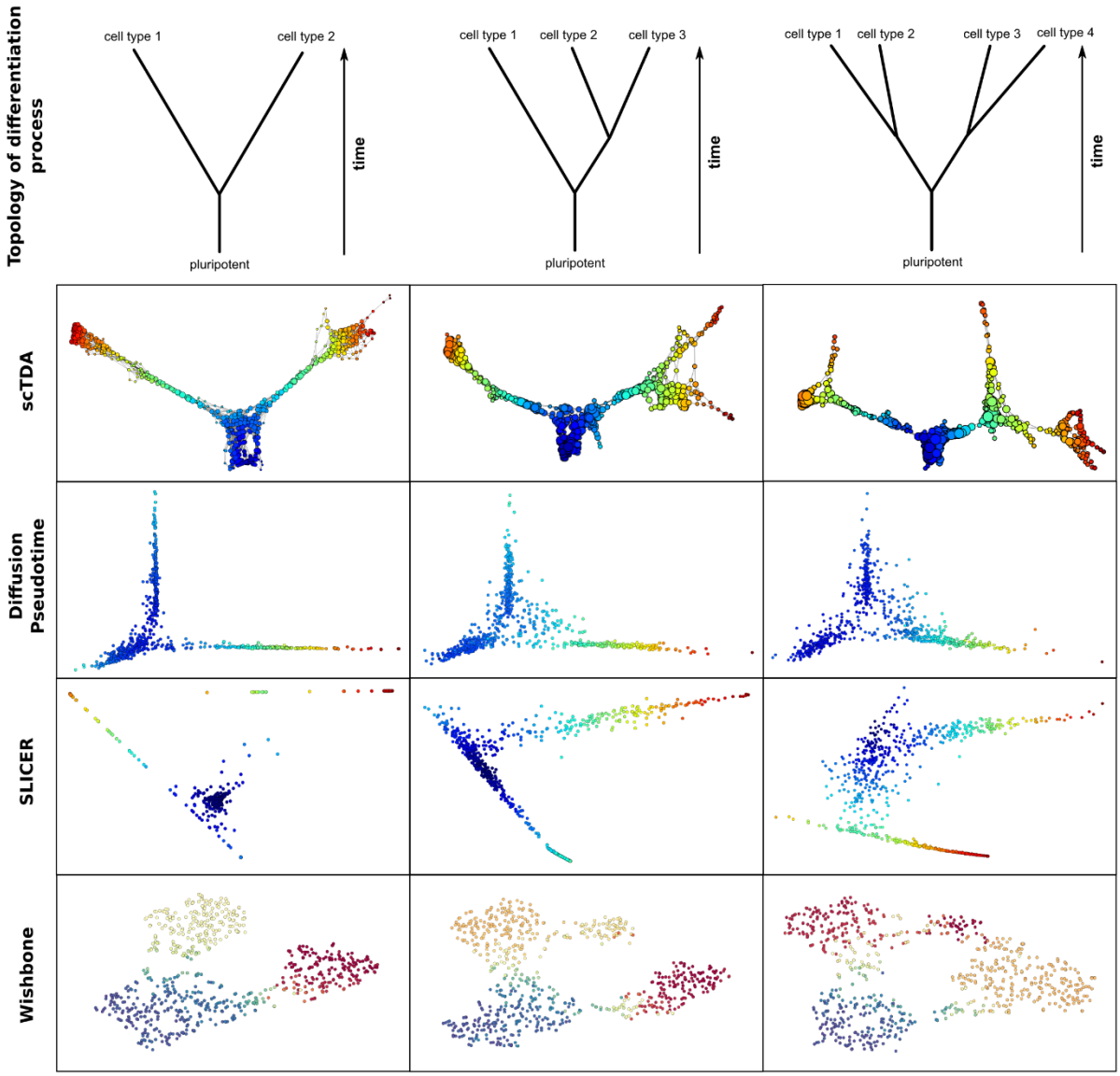
Supplementary Figures



Supplementary Figure 1. Workflow for single cell Topological Data Analysis (scTDA).

Longitudinal single-cell RNA-Seq data is mapped, demutiplexed and pre-processed, which removes cells that do not pass stringent quality controls (QC). A topological representation is built using the Mapper algorithm, and is based on highly-expressed genes that have a large variability. The expression of each transcript is a function that supports the topological representation. A pseudo-time ordering is established and different statistics (gene connectivity, centroid, dispersion) are computed, making it possible to

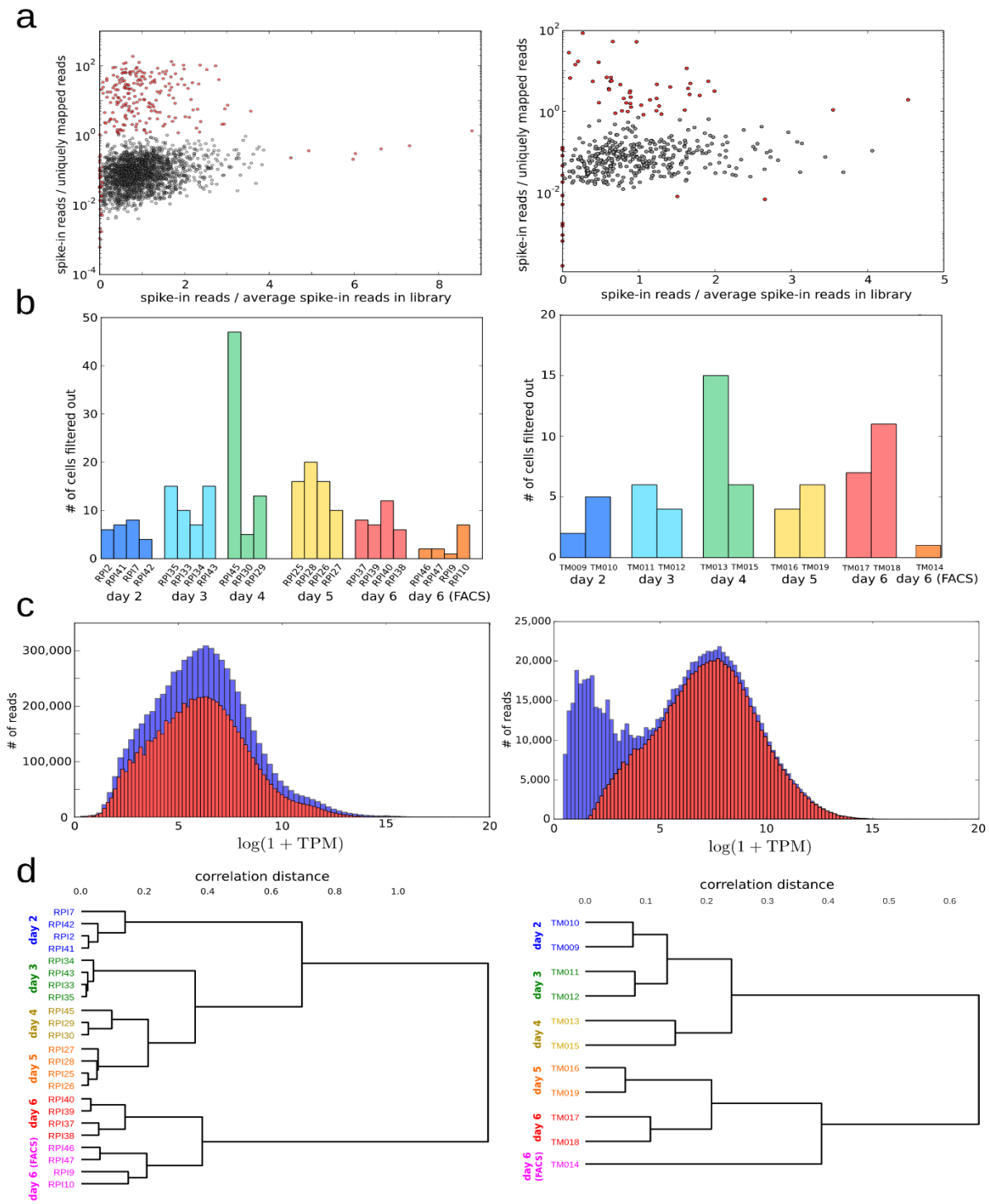
identify expression programs associated with different stages and populations of cells generated throughout the differentiation process.



Supplementary Figure 2. Comparison of various algorithms for the analysis of single cell expression data from developmental processes, using simulated data.

Noisy, branched, asynchronous differentiation processes were simulated, including the effect of drop-out events. 700 cells were sampled at three time points. The reconstruction of the differentiation process produced by scTDA, Differentiation Pseudotime, Slicer, and Wishbone is shown, colored by the inferred

pseudotime, for processes with one, two, or three branching points. scTDA reproduces the topology of the differentiation process and its temporal structure in complex situations with more than one branching point.



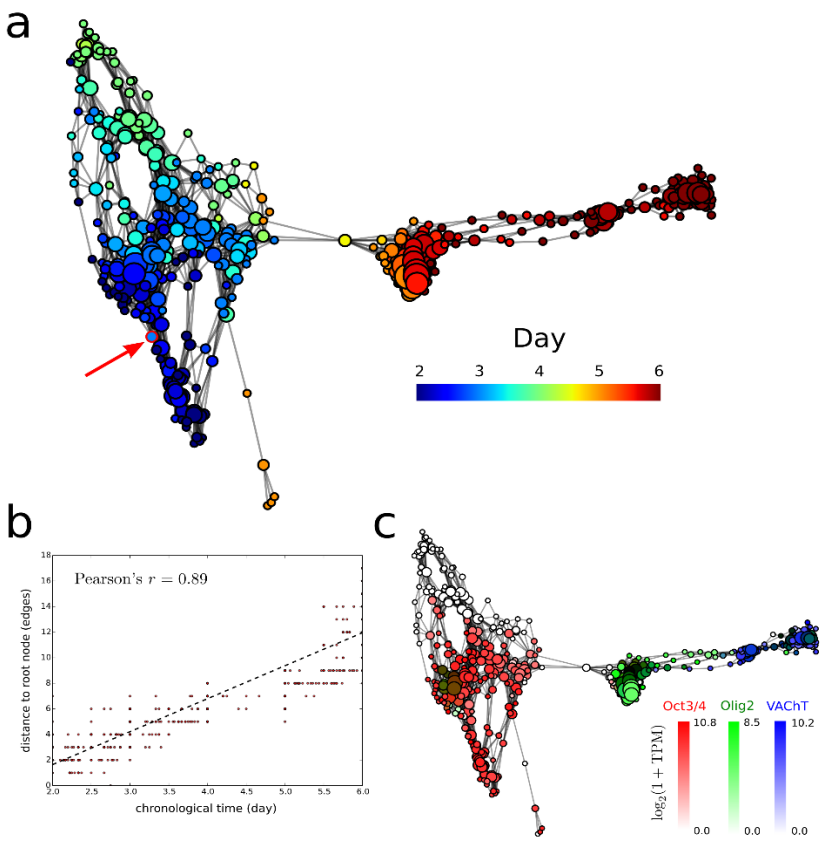
Supplementary Figure 3. Single-cell RNA-seq data filtering in the main (left) and pilot (right) motor neuron differentiation experiments.

a. The fraction of reads from each cell, represented by ERCCs (the ratio between ERCC reads and uniquely mapped reads), was plotted against the relative level of ERCCs per cell (the ratio between ERCC reads and the average number of ERCC reads in the library). This was done for each of the 2,208 cells sequenced in the main experiment (one library demonstrating a large batch effect was removed from the analysis), and the 440 cells in the pilot experiment. Cells that are filtered out are represented in red. The ordinate is represented in logarithmic scale.

b. The distribution of filtered-out cells across each of the libraries.

c. The distribution of expression values for all genes in the complete set of cells before (blue) and after (red) filtering. In the pilot experiment (right), the expression of genes supported by less than 5 reads in a cell was set to zero to remove noise near the mRNA capture threshold.

d. Hierarchical clustering of the expression centroids of each library, based on correlation distance. Libraries cluster according to day, with substantial overlap between libraries from the same day.

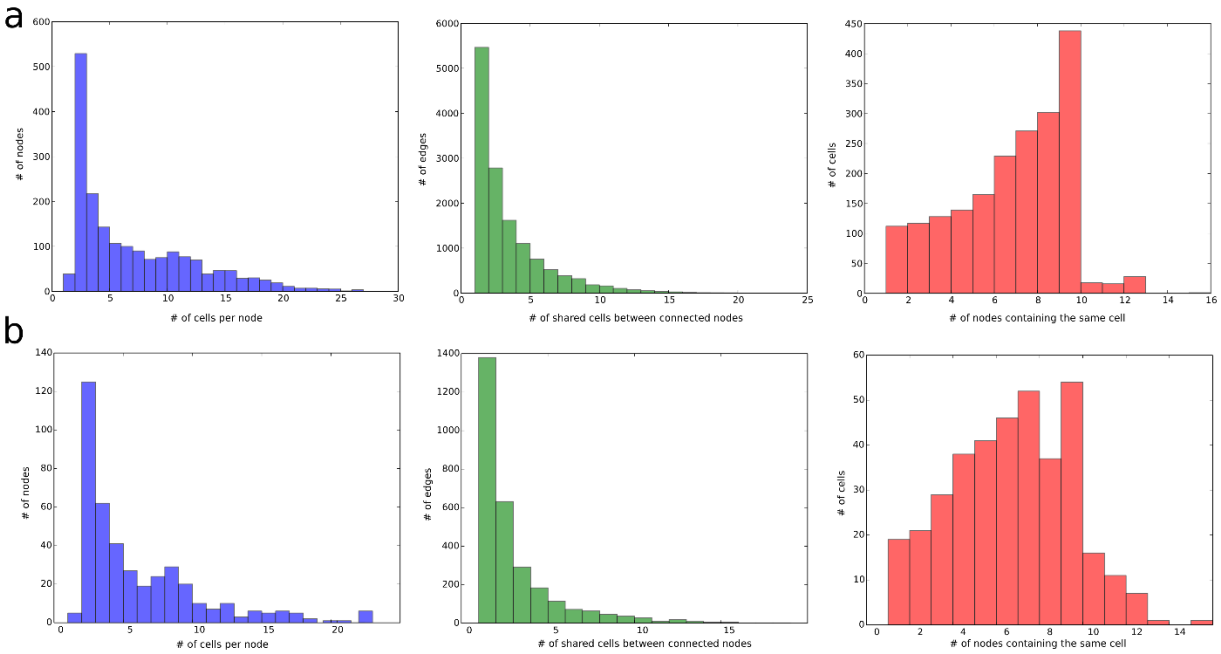


Supplementary Figure 4. Topological representation of longitudinal single-cell RNA-seq data from mESC differentiation into motor neurons in the pilot experiment.

a. The topological representation labeled by sampling time.

b. The distance of each node to the root node (marked with a red arrow in **a**), represented as a function of sampling time. The chronological time of a node is defined as the mean of the sampling times of the cells in the node.

c. The topological representation labeled by mRNA levels of known markers of pluripotent cells (*Oct3/4*, red), motor neuron progenitors (*Olig2*, green) and post-mitotic neurons (*VAChT*, blue).

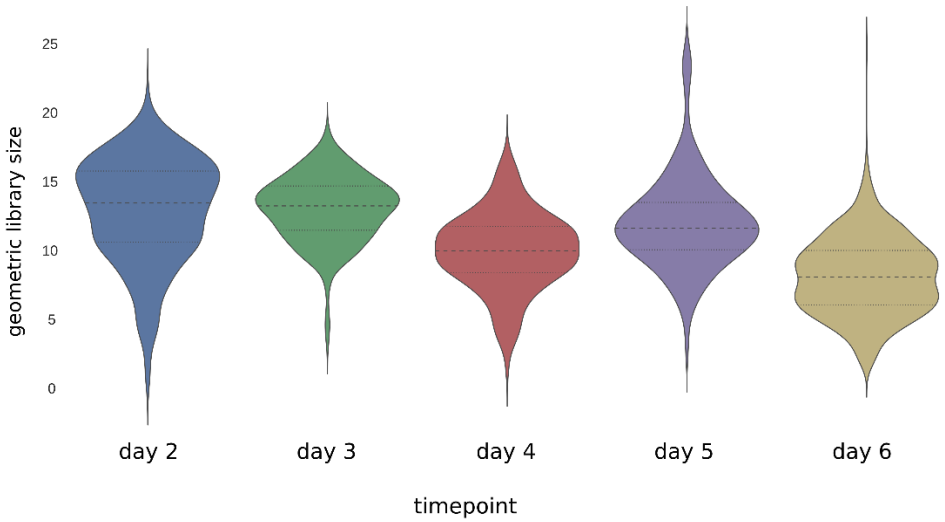


Supplementary Figure 5. Statistics of the topological representation in the main (a) and pilot (b) experiments.

Left: The distribution of the number of cells associated to each node of the topological representation.

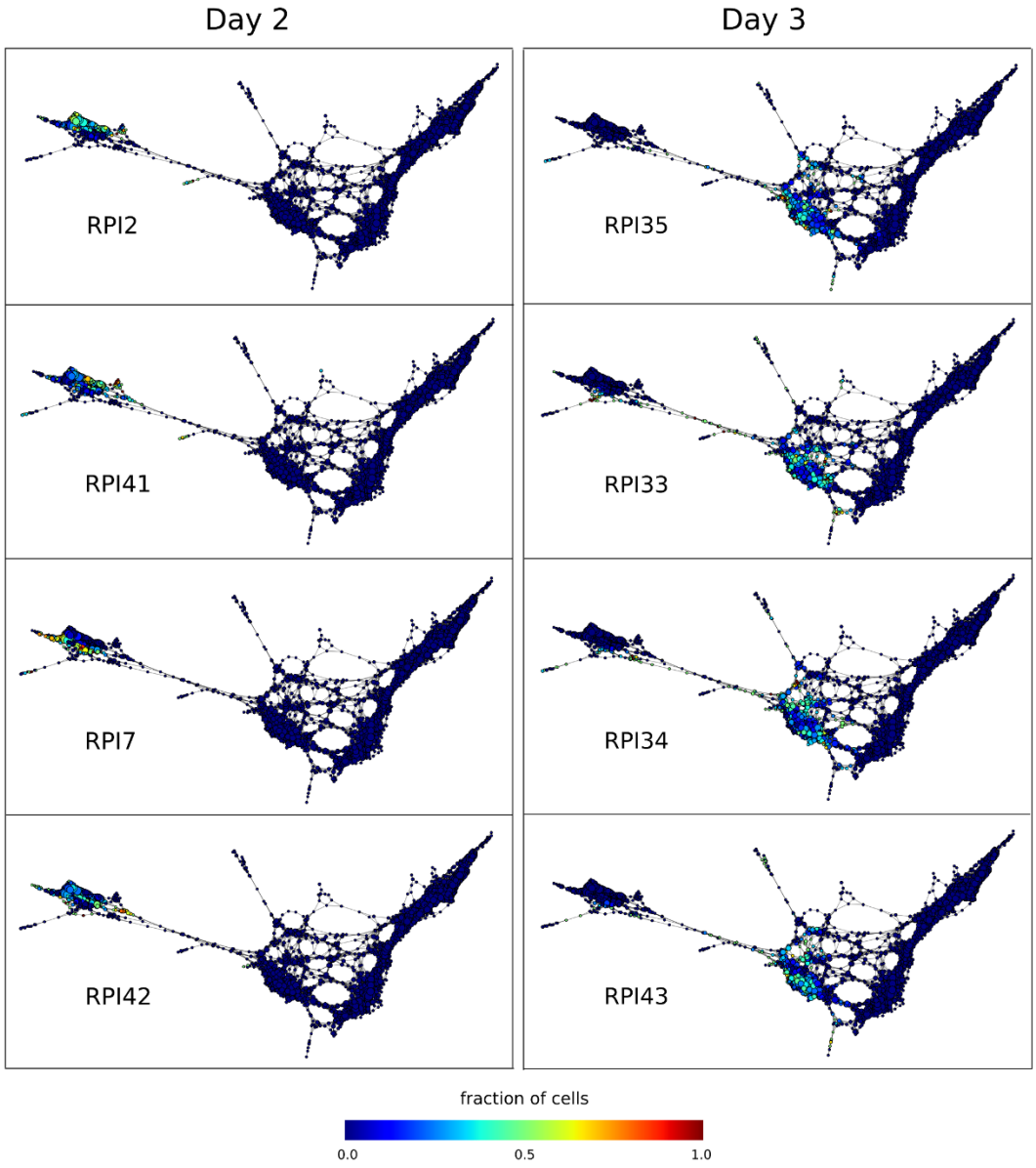
Middle: The distribution of the number of cells that are shared between nodes connected by an edge.

Right: The distribution of the number of nodes in which an individual cell appears.



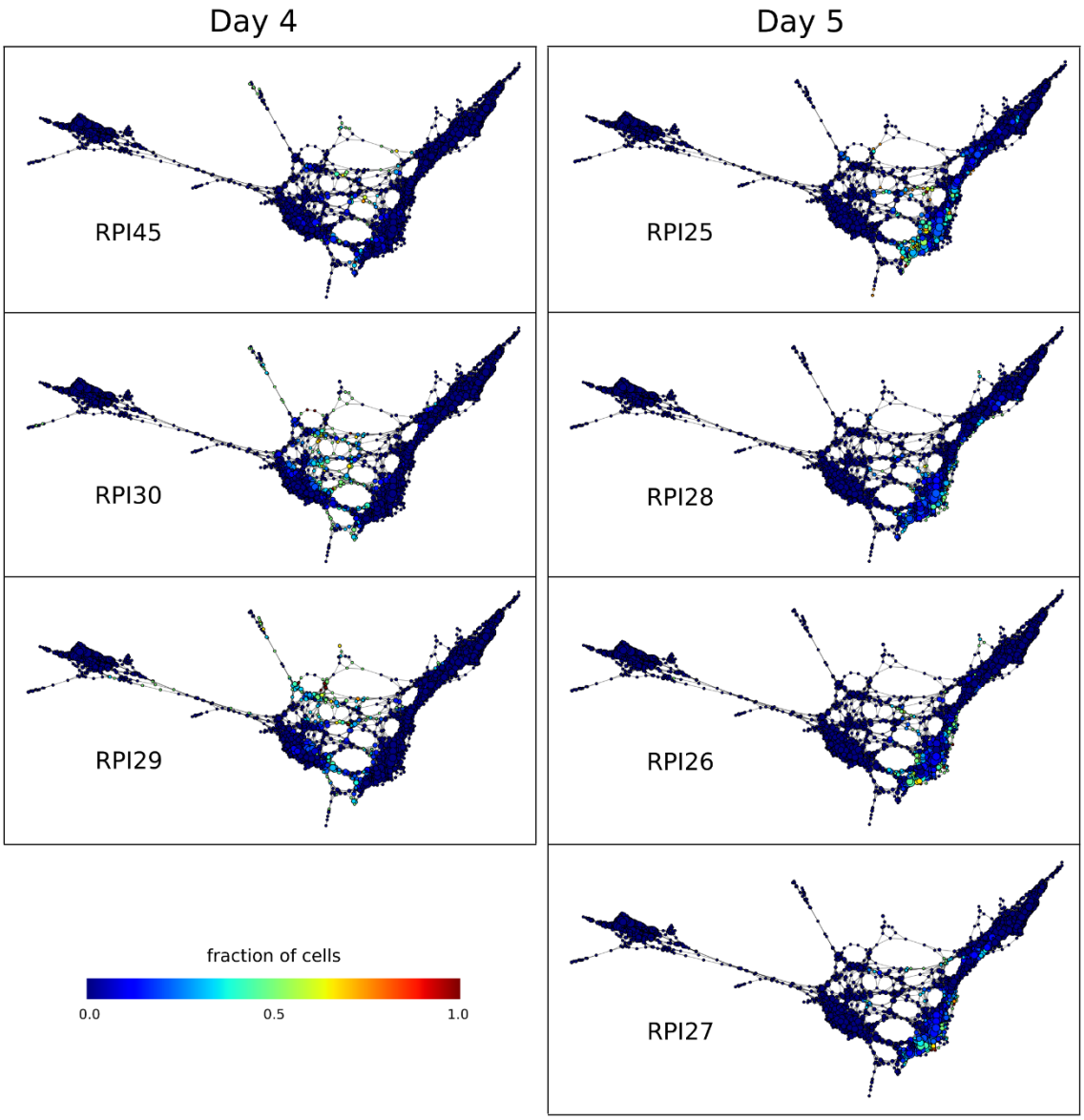
Supplementary Figure 6. Dependence of the library complexity on the sampling time in the main motor neuron differentiation experiment.

The distribution of the geometric library size is plotted as a function of the sampling time, showing a mild dependence between these two quantities.



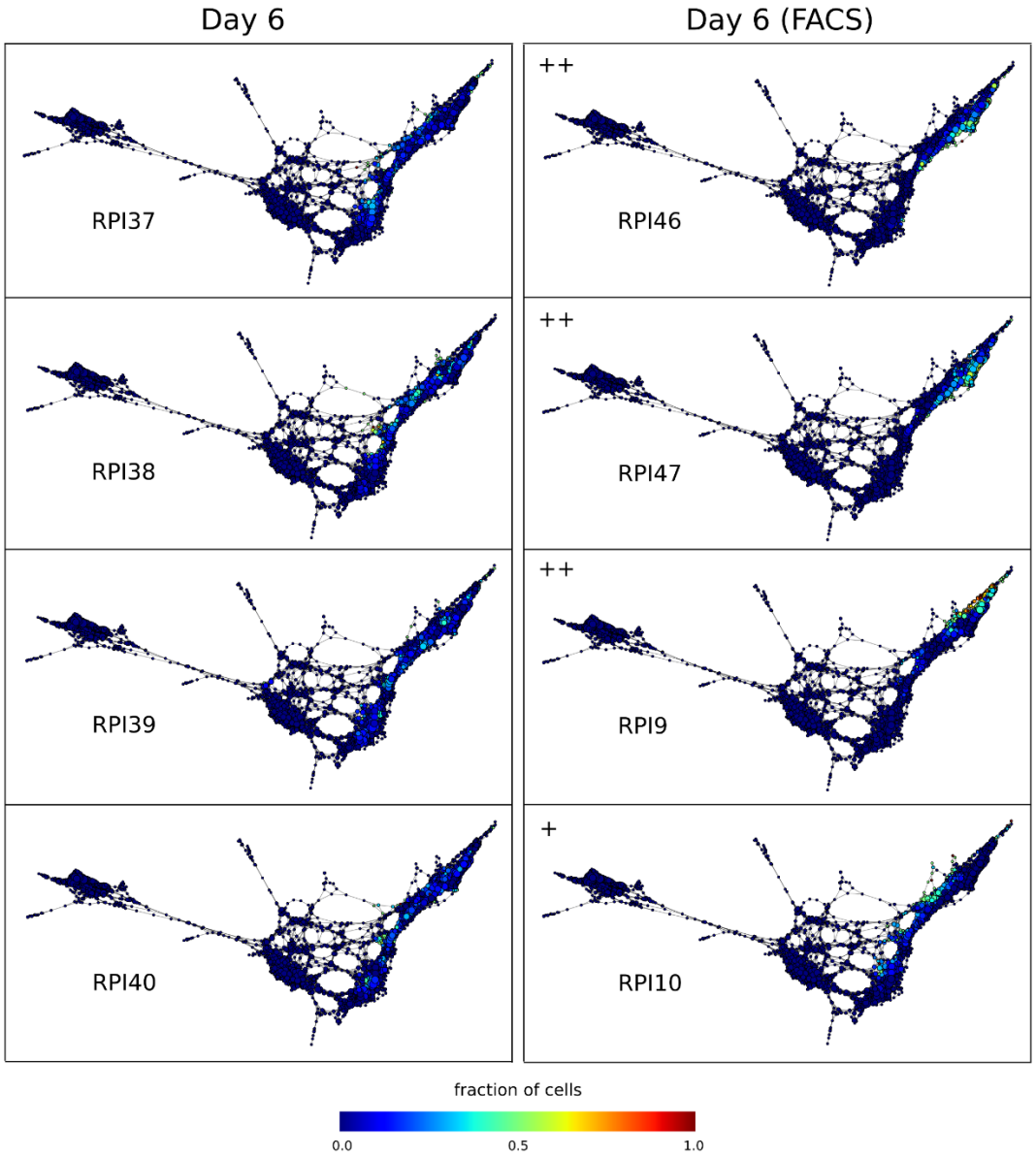
Supplementary Figure 7 (1/3). Distribution of different sequencing libraries across the topological representation in the main experiment.

Libraries from the same day have a substantial overlap in the topological representation, reflecting the absence of large batch effects.



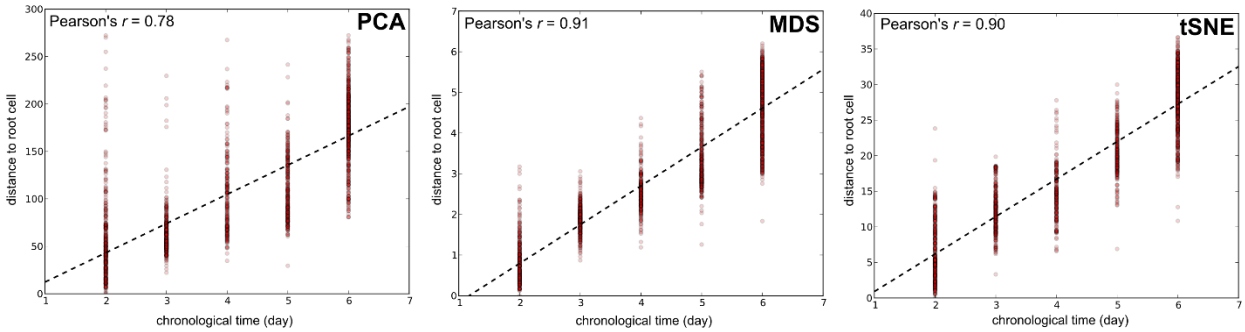
Supplementary Figure 7 (cont., 2/3). Distribution of different sequencing libraries across the topological representation in the main experiment.

Libraries from the same day have a substantial overlap in the topological representation, reflecting the absence of large batch effects.



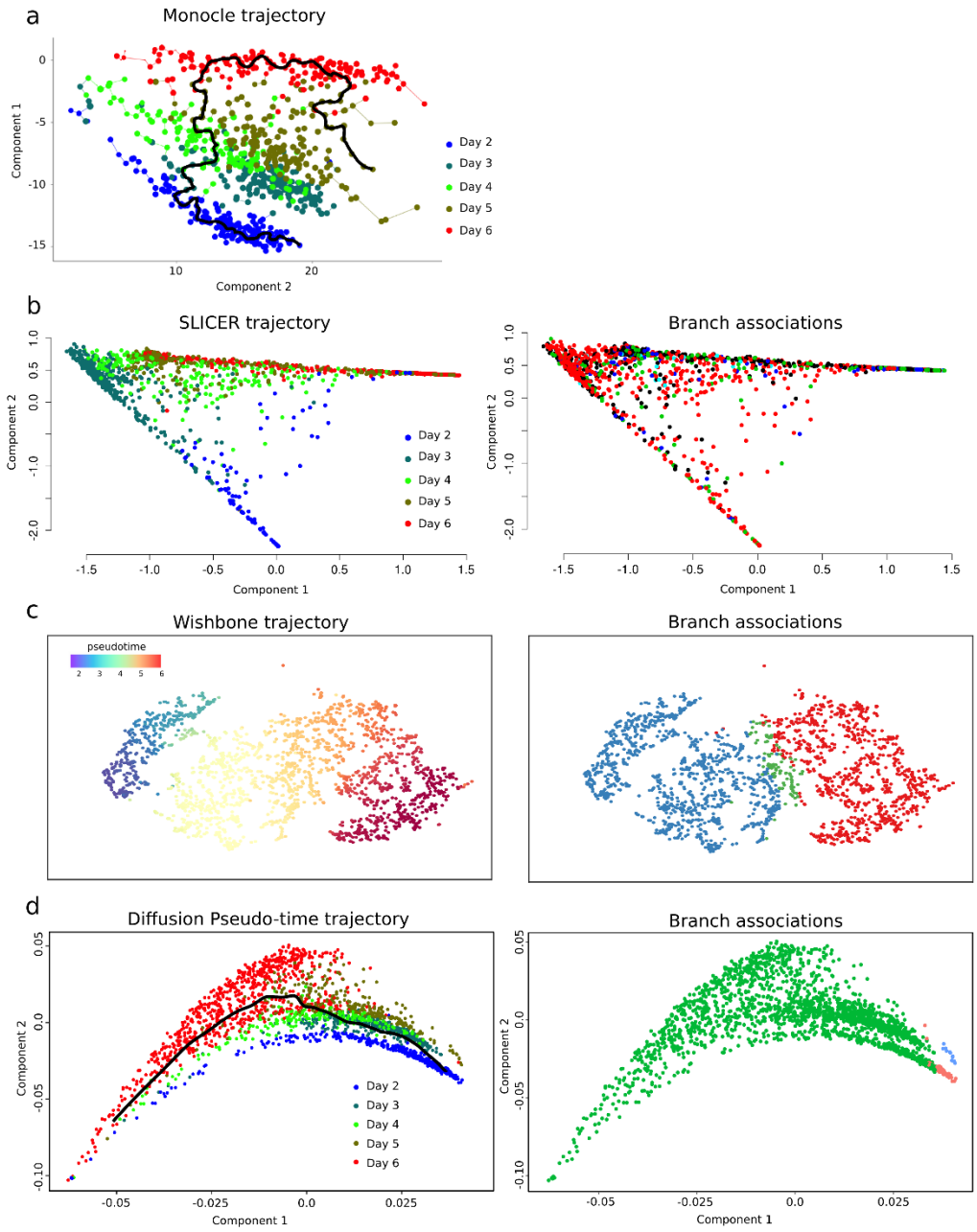
Supplementary Figure 7 (cont., 3/3). Distribution of different sequencing libraries across the topological representation in the main experiment.

Libraries from the same day have a substantial overlap in the topological representation, reflecting the absence of large batch effects.



Supplementary Figure 8. Correlation between distance to root node and chronological time for PCA, MDS, and tSNE representations of the single cell expression data from mESC differentiation into motor neurons.

The two-dimensional Euclidean distance to the root cell (defined as the one that maximizes the correlation with the chronological sampling time) is shown as a function of the chronological sampling time.



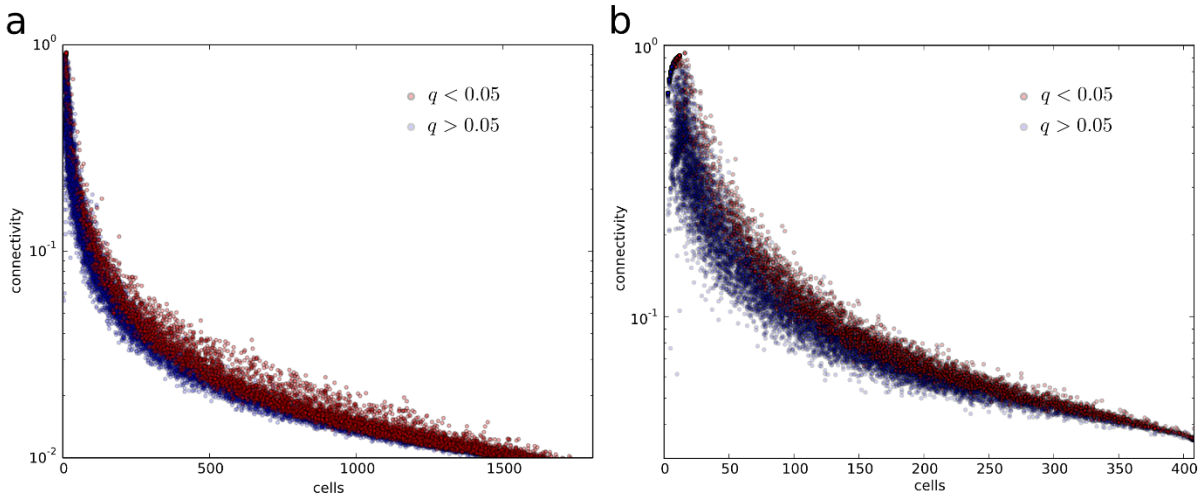
Supplementary Figure 9. Comparison to other algorithms for the analysis of longitudinal single-cell RNA-seq experiments, using data from mESC differentiation into motor neurons.

a. Output produced by Monocle. Monocle was run on 834 cells from the main experiment to build a two dimensional minimum spanning tree representation of the data. Cells are colored by sampling day.

b. Output produced by SLICER. SLICER was run on the 1,964 cells from the main experiment to build a two dimensional representation of the data, where cells are assigned a branch in the differentiation tree inferred from the data. Left: Cells are colored by sampling day. Right: Cells are colored according to the branch assignments made by SLICER.

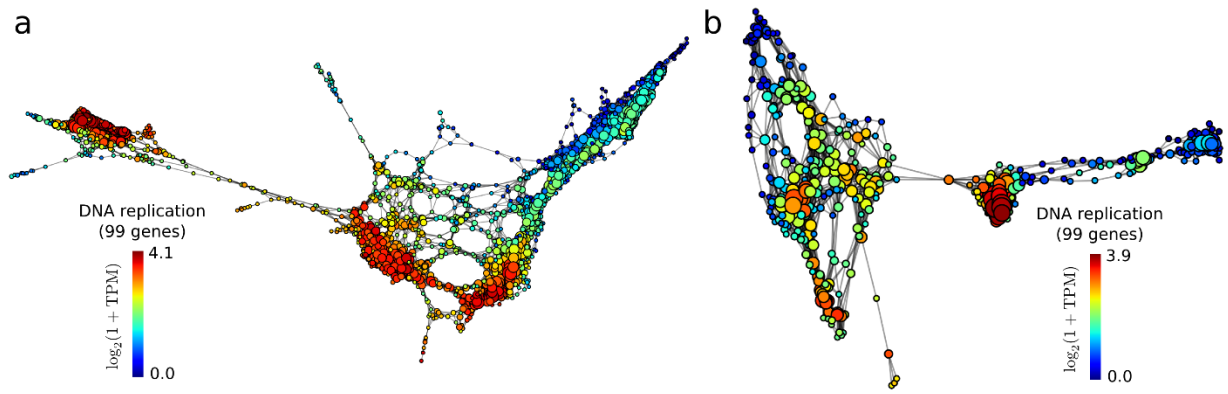
c. Output produced by Wishbone. Wishbone was run on the 1,964 cells from the main experiment to build a two dimensional t-SNE representation of the data where cells are assigned a pseudo-time and a branch in the differentiation structure inferred from the data. Left: Cells are colored by pseudo-time. Right: Cells are colored according to the branch assignments made by Wishbone.

d. Output produced by Diffusion Pseudotime. Diffusion Pseudotime was run on the 1,964 cells from the main experiment to build a two dimensional representation of the data based on the first two diffusion components, where cells are assigned a pseudo-time and a branch in the differentiation structure inferred from the data. Left: Cells are colored by pseudo-time. Right: Cells are colored according to the branch assignments made by Diffusion Pseudotime.



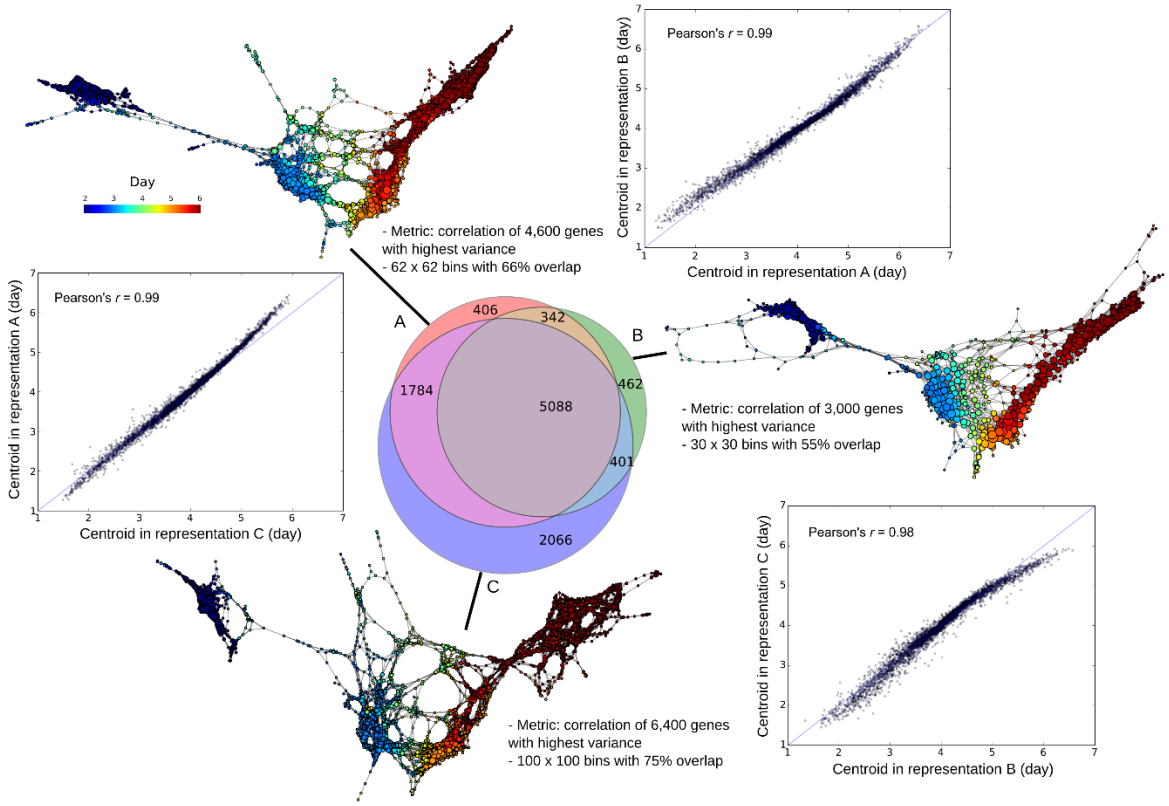
Supplementary Figure 10. Distribution of the gene connectivity score against the number of cells expressing the gene in the main (left) and pilot (right) motor neuron differentiation experiments.

Statistical significance was evaluated by means of a permutation test (Online Methods). Genes with a significant gene connectivity (q < 0.05) after controlling for multiple hypothesis testing are shown in red.



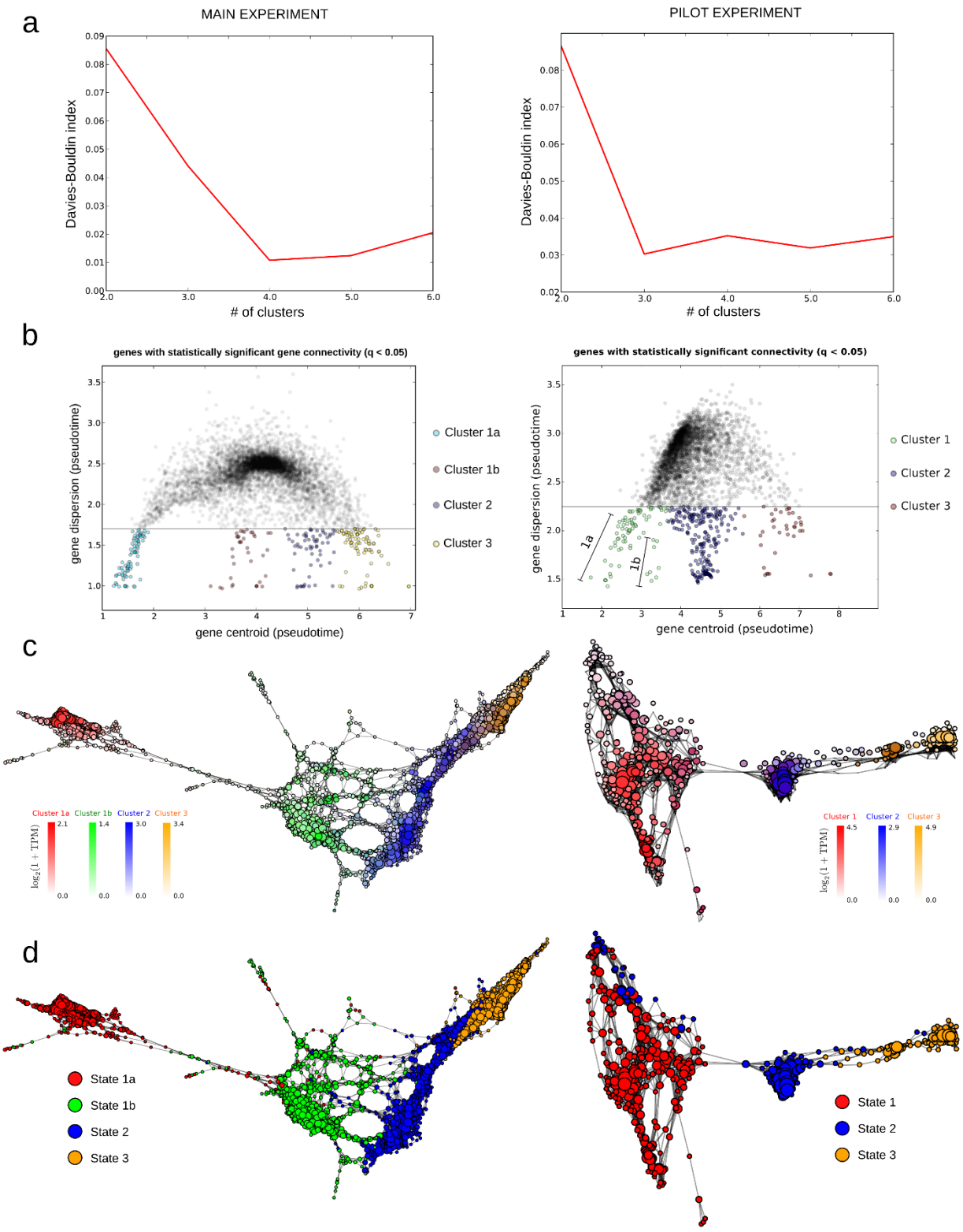
Supplementary Figure 11. Expression levels of genes associated to DNA replication in the main (a) and pilot (b) motor neuron differentiation experiments.

Topological representation labeled according to the expression of 99 genes related to DNA replication. Differentiated neurons exhibit very low levels of DNA replication, consistent with post-mitotic arrest.



Supplementary Figure 12. Stability of topological representations against different parameter choices.

Topological representations of the single cell RNA-seq data of the main motor neuron differentiation experiment using different choices for the number of patches and their overlap considered in the Mapper algorithm, as well as for the number of genes that are used to compute the distance matrix. The Venn diagram displays the number of genes with significant gene connectivity ($q < 0.05$) in three topological representations, showing a large degree of consistency across different representations. For each pair of representations, the correlation between the centroids (expressed in pseudo-time) of the $n = 5{,}088$ genes with significant ($q < 0.05$) connectivity in all representations is also shown. The distribution of gene centroids is highly consistent across different choices of parameters, with Pearson's correlations in the range 0.98 – 0.99.


a
MAIN EXPERIMENT
PILOT EXPERIMENT
Davies-Bouldin index
# of clusters
b
genes with statistically significant gene connectivity (q < 0.05)
genes with statistically significant connectivity (q < 0.05)
gene dispersion (pseudotime)
gene centroid (pseudotime)
Cluster 1a
Cluster 1b
Cluster 2
Cluster 3
Cluster 1
1a
1b
c
d
State 1a
State 1b
State 2
State 3
State 1
State 2
State 3

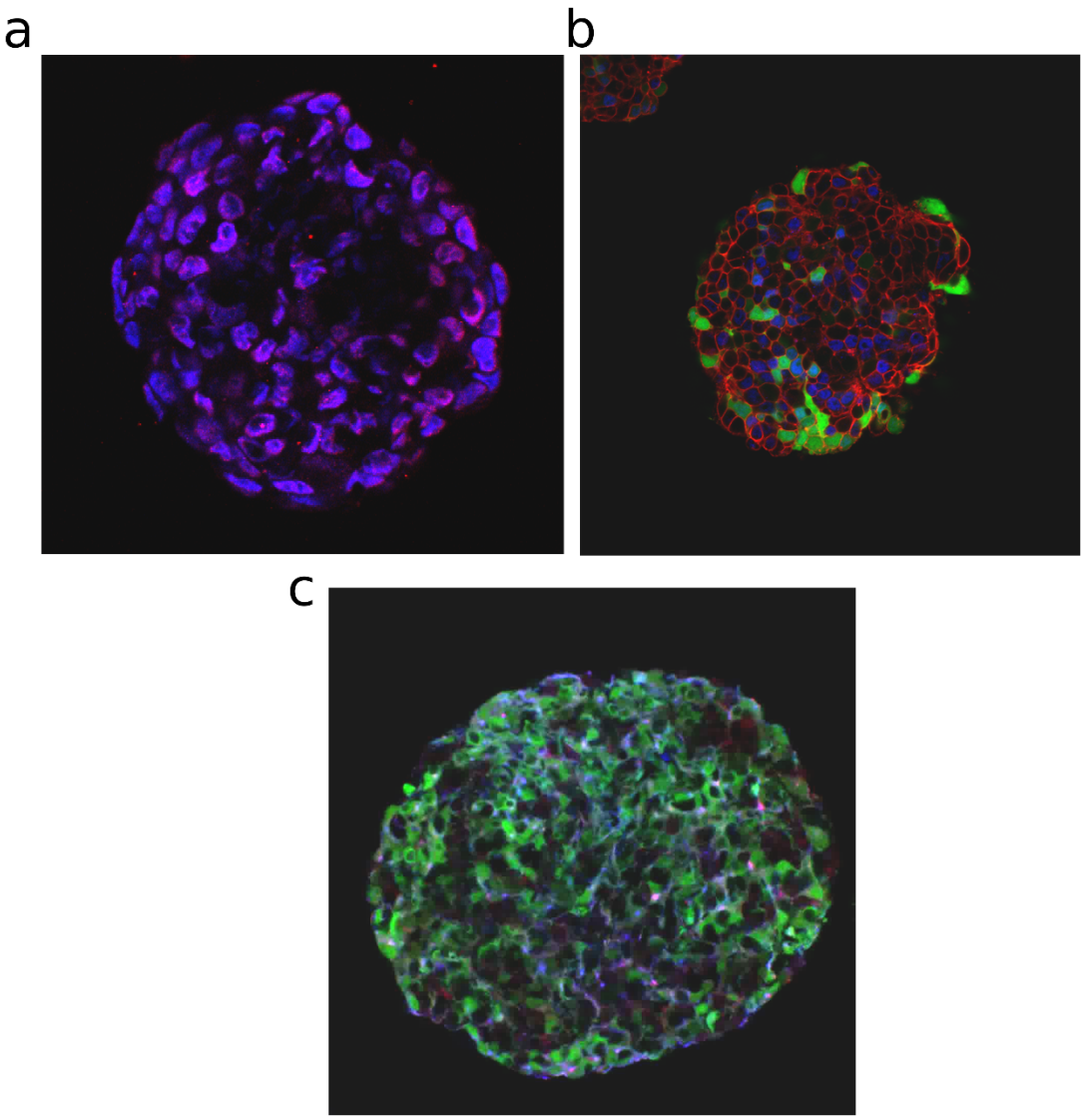
Supplementary Figure 13. Cellular populations arising throughout the differentiation of mESCs into motor neurons in the main (left) and pilot (right) experiments.

a. Davies-Bouldin index for k-means clustering of the centroid of low-dispersion genes in the topological representation of the main experiment (k_i < 1.7) and the pilot experiment (k_i < 2.25). The minimum is achieved for four and three clusters, respectively, in the main and pilot experiments.

b. Distribution of the centroid and dispersion of significant genes (q < 0.05) in the topological representation of the main experiment. Four principal gene groups (1a, 1b, 2 and 3) naturally arise from clustering the centroids of low-dispersion genes, corresponding to four transient cellular populations arising throughout the differentiation.

c. Topological representation of the two experiments labeled by mRNA levels of genes in each of the gene clusters obtained by k-means.

d. Topological representation of the two experiments labeled by the state assigned to each node based on the expression levels of each of the gene clusters.

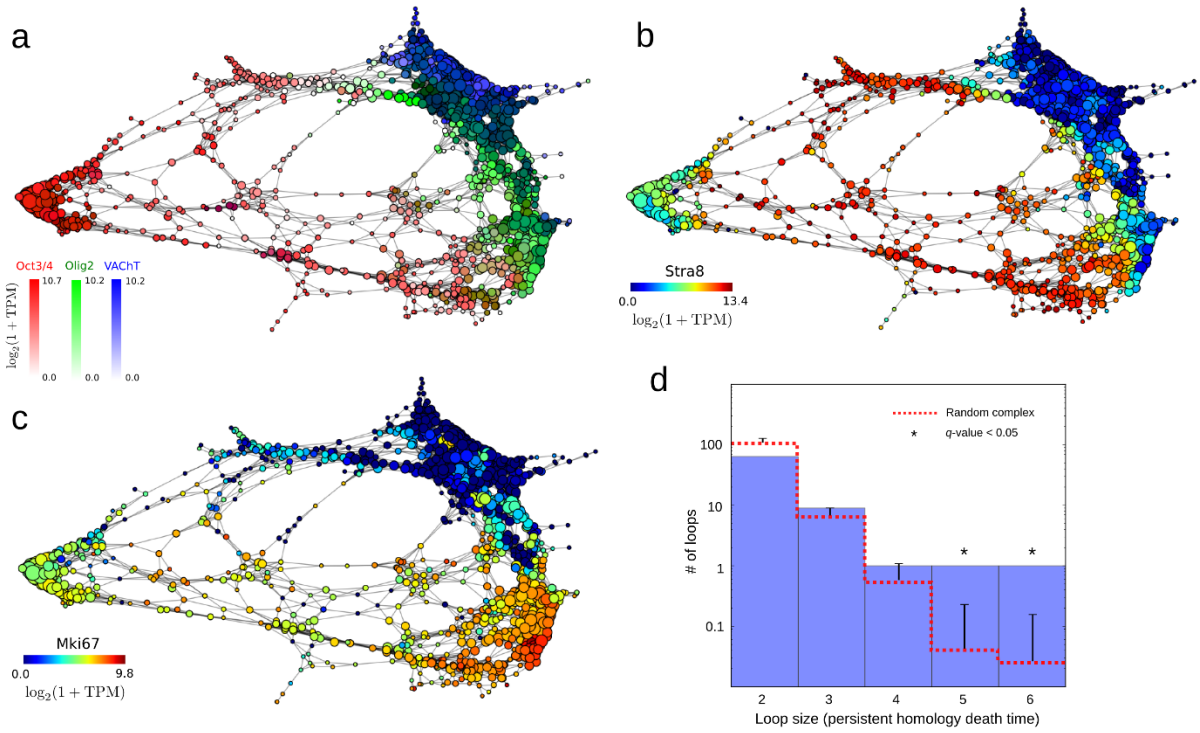


Supplementary Figure 14. Immunostaining of murine EBs against stage-specific surface markers identified by scTDA.

a. Immunostaining against the pluripotent surface marker *Pecam1*. Immunostaining of day 2 EBs against *Pecam1* (red), *Oct3/4* (blue), and *MNx1::eGFP* (green), showing overlap between *Pecam1*$^{+}$ cells and *Oct3/4*$^{+}$ cells.

b. Immunostaining against the neural progenitor surface marker *Ednrb*. Immunostaining of day 5 EBs against *Ednrb* (red), *Olig2* (blue), and *Mnx1::eGFP* (green), showing partial overlap between *Ednrb*$^{+}$ cells and *Olig2*$^{+}$ cells, and mutual exclusivity between *Ednrb*$^{+}$ cells and *Mnx1::eGFP*$^{+}$ cells.

c. Immunostaining against the post-mitotic neuron surface marker *Slc10a4*. Immunostaining of day 6 EBs against *VAChT* (red), *Slc10a4* (blue), and *Mnx1::eGFP* (green), showing overlap between *VAChT*$^{+}$ cells and *Slc10a4*$^{+}$ cells, and partial overlap with *Mnx1::eGFP*$^{+}$ cells.



Supplementary Figure 15. Topological representation of single cell expression data from mESC differentiation into motor neurons, based on cell cycle genes only.

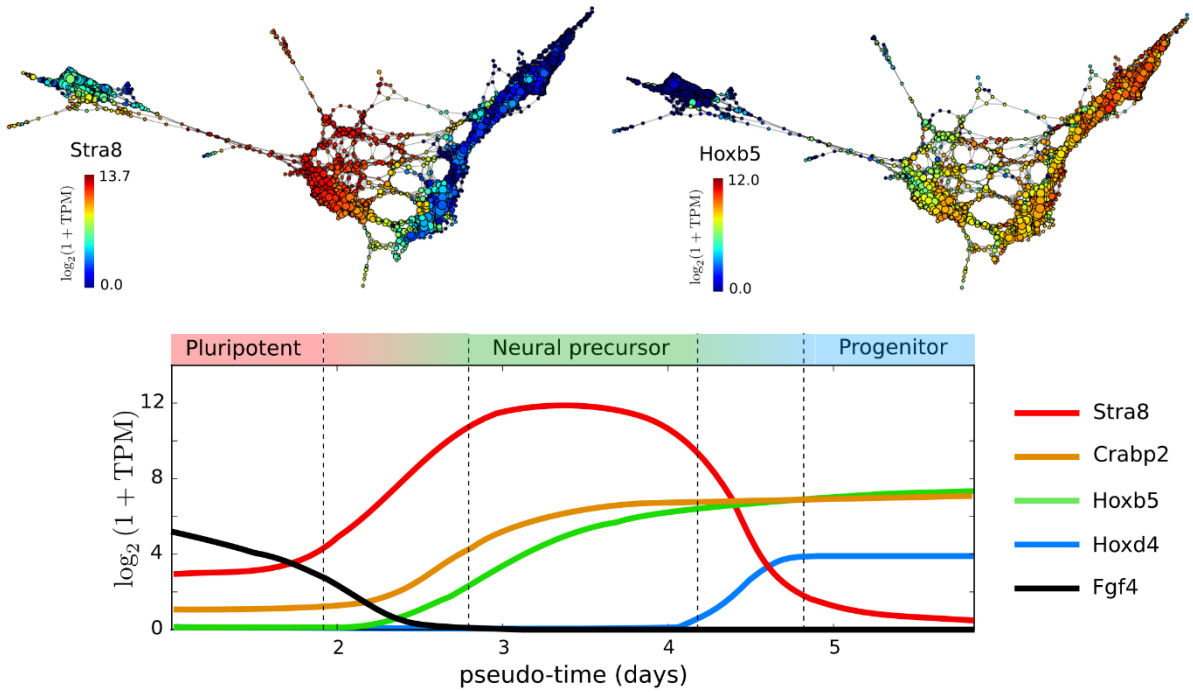
a. The topological representation is labeled by mRNA levels of known markers of pluripotent cells (*Oct3/4*, red), motor neuron progenitors (*Olig2*, green) and post-mitotic neurons (*VAChT*, blue).

b. The region of the topological representation corresponding to neural precursors has numerous loops. The topological representation is labeled by mRNA levels of *Stra8*, a marker of neural precursors.

c. Neural precursors separate into proliferative and non-proliferative populations. The topological representation is labeled by *Mki67*, a known marker of cellular proliferation.

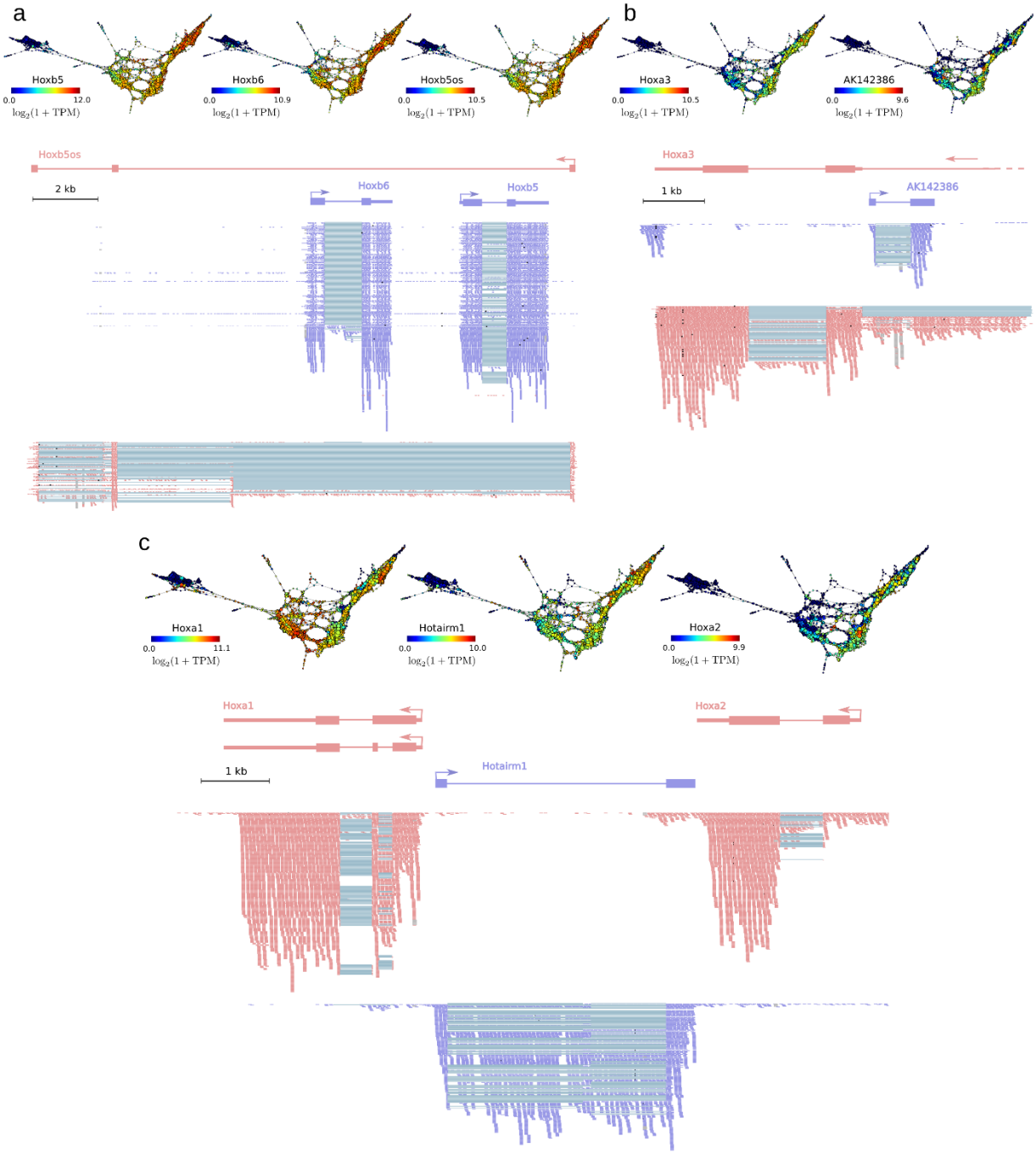
d. Distribution of the number of loops of a given size (as measured by their first persistent homology death time) across the representation (blue), compared to that of random complexes built by permuting

the genes independently in each cell. Some of the larger loops in the topological representation are statistically significant, consistent with a biological origin for these loops.



Supplementary Figure 16. Transcriptional events occurring during the transition between pluripotent and neural precursor, and between neural precursor and neural progenitor populations of cells.

Reconstructed expression timeline for some of the transcriptional changes identified by scTDA in the transitions between pluripotent and neural precursor, and between neural precursor and neural progenitor populations, in the main motor neuron differentiation experiment. scTDA identified upregulation of *Stra8* and downregulation of *Fgf4* as one of the early events in the transition between a pluripotent and a neural precursor state. The topological representation labeled according to the mRNA expression levels of *Stra8* and *Hoxb5* is also shown.

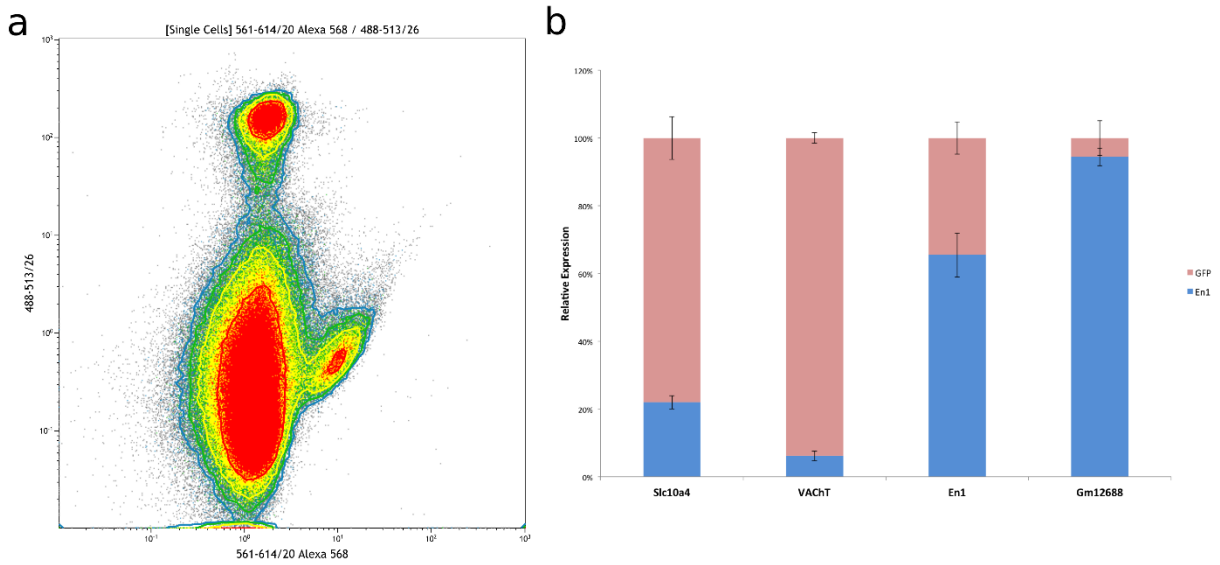

a
Hoxb5
0.0
12.0
log2(1 + TPM)
Hoxb6
0.0
10.9
log2(1 + TPM)
Hoxb5os
0.0
10.5
log2(1 + TPM)
Hoxb5os
2 kb
Hoxb6
Hoxb5
b
Hoxa3
0.0
10.5
log2(1 + TPM)
AK142386
0.0
9.6
log2(1 + TPM)
Hoxa3
1 kb
AK142386
c
Hoxa1
0.0
11.1
log2(1 + TPM)
Hotairm1
0.0
10.0
log2(1 + TPM)
Hoxa2
0.0
9.9
log2(1 + TPM)
Hoxa1
Hoxa2
1 kb
Hotairm1

Supplementary Figure 17. Expression of antisense lncRNAs derived from homeobox gene clusters during the differentiation of mESCs into motor neurons.

a. Topological representations of the main experiment labeled by mRNA levels of *Hoxb5*, *Hoxb6*, and the antisense lncRNA *Hoxb5os*, showing concordant expression of these genes during the generation of motor neurons from mESCs. Stranded bulk RNA-seq reads from day 5 of the differentiation process are also depicted.

b. Topological representations of the main experiment labeled by mRNA levels of *Hoxa3* and the antisense lncRNA *AK142386,* and stranded bulk RNA-seq reads.

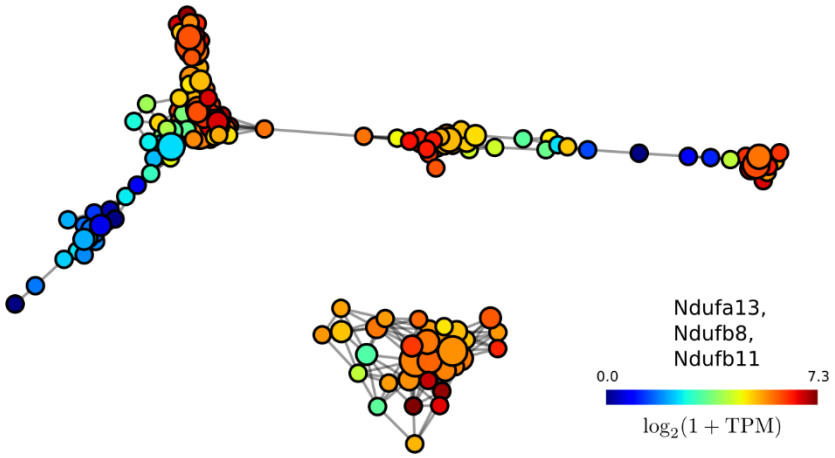
c. Topological representations of the main experiment labeled by mRNA levels of *Hoxa1*, *Hoxa2*, and the antisense lncRNA *Hotairm1*, and stranded bulk RNA-seq reads.



Supplementary Figure 18. Sorting of V1 interneurons validates the exclusive expression of lncRNA *Gm12688*.

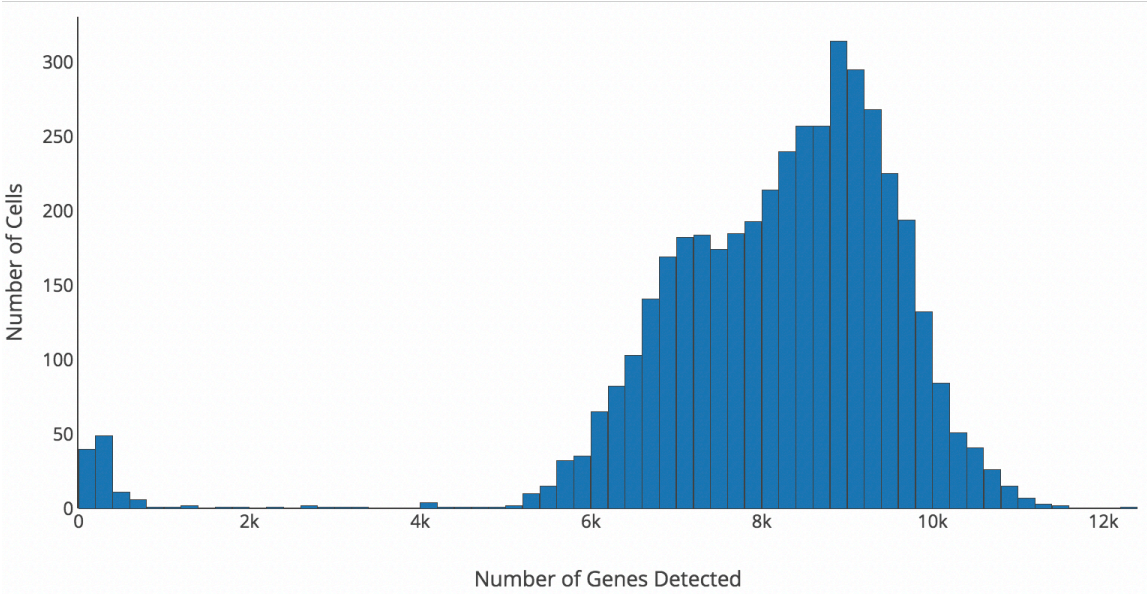
a. Dissociation of day 6 EBs and immunostaining against Engrailed (*En1*) enabled purification of V1 interneurons through flow cytometry. Alexa 568 expressing cells are *En1* positive. *Mnx1::eGFP* expressing motor neurons are low in Alexa 568 signal, but high in eGFP fluorescence.

b. RT-PCR of eGFP expressing motor neurons and *En1*$^{+}$ cells demonstrates enrichment of expected markers, and a high enrichment of *Gm12688* in V1 interneurons.



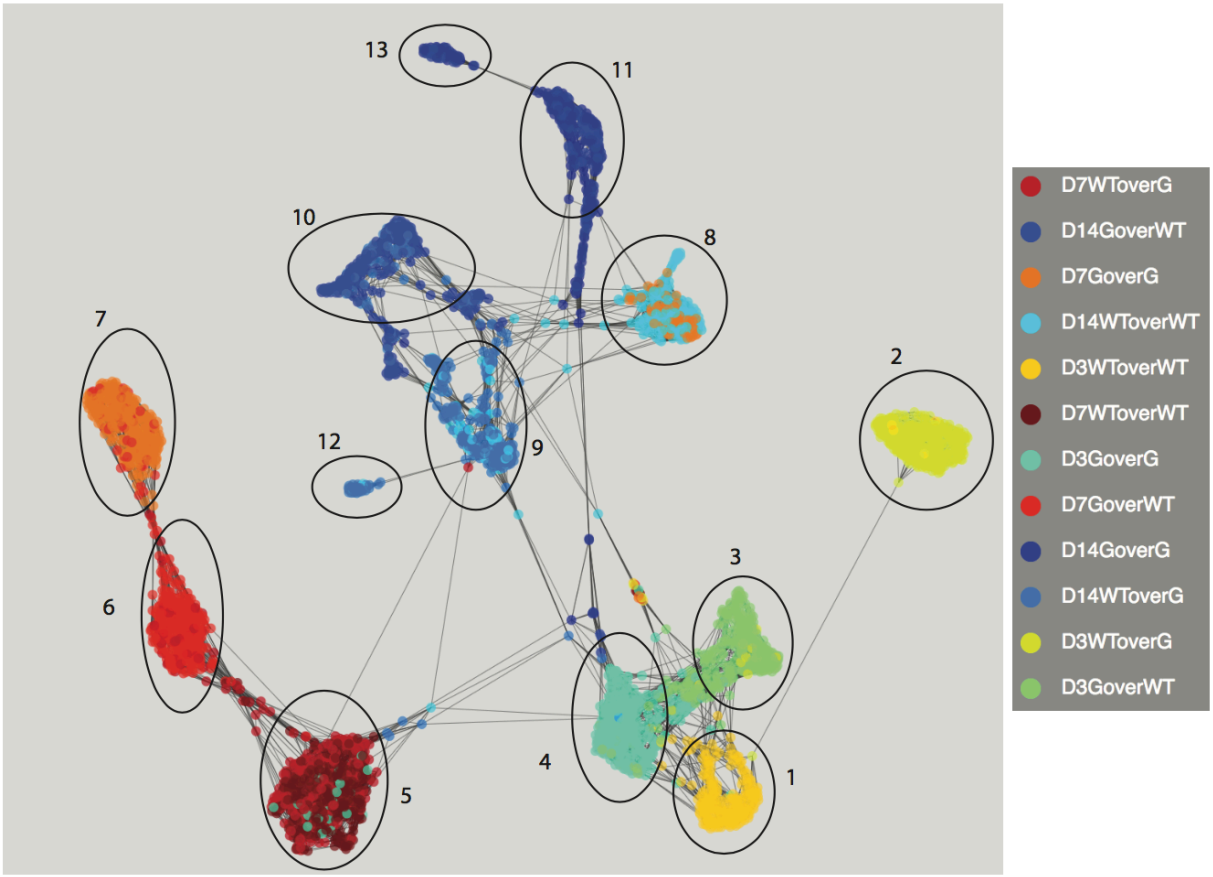
Supplementary Figure 19. Differential expression of genes across the two populations of alveolar type I cells in the developing distal lung epithelium.

The topological representation is labeled by the mRNA levels of genes coding for proteins from the mitochondrial respiratory chain NADH dehydrogenase that are differentially expressed between the two populations of alveolar type I cells identified by scTDA.



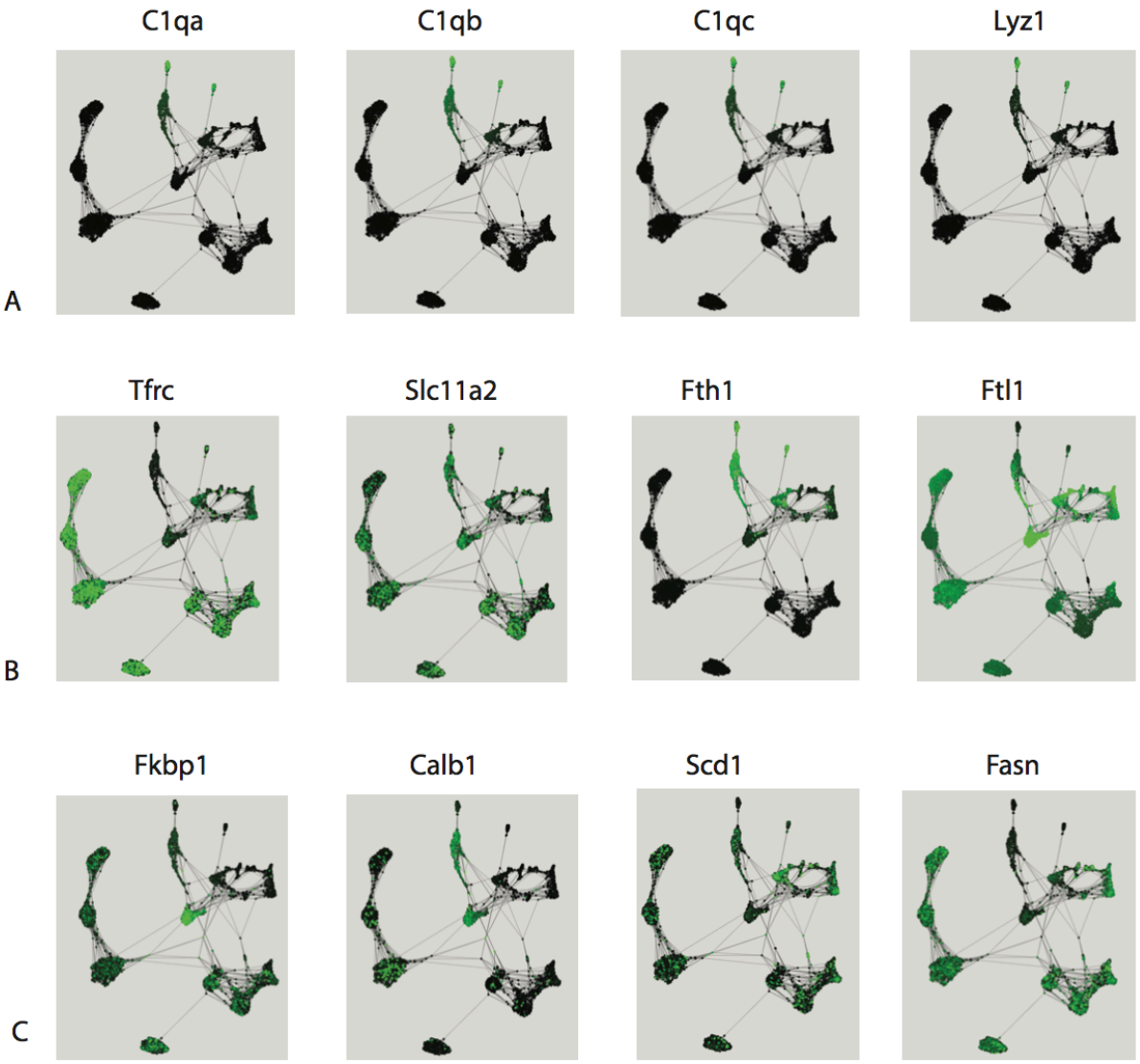
Supplementary Figure 20. Histogram of numbed of genes detected per cell.

An average of 8500 genes were detected per cell of the ESMN timecourse. Cells that had less than 5000 genes detected were discarded from the analysis.



Supplementary Figure 21: Clusters of ESMNs in kNN graph

Clusters are identified based on cell density and labeled for reference.

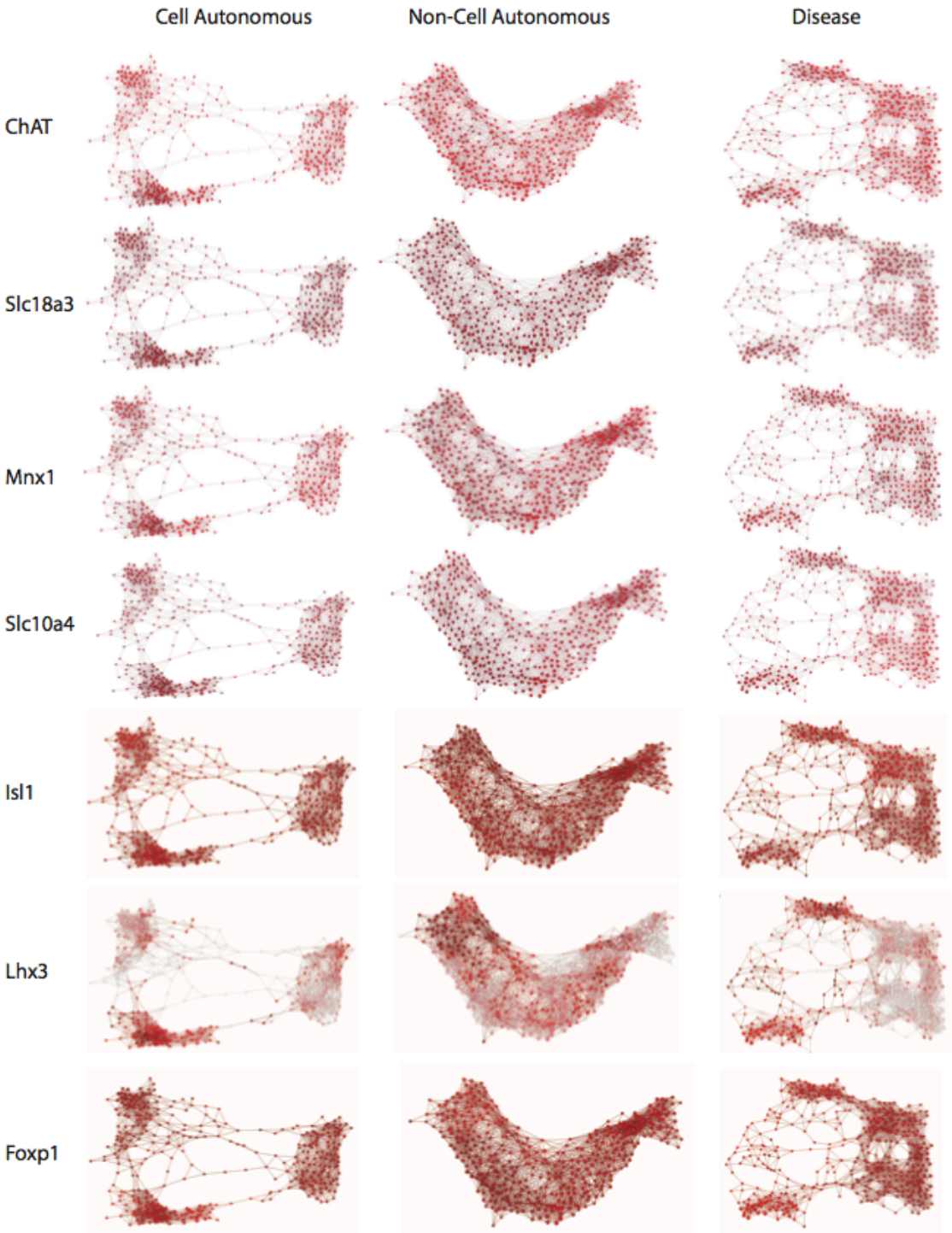


Supplementary Figure 22: Gene expression in kNN graph

A) Apoptotic markers define clusters 12 and 13

B) Iron responsive genes indicate disbalance in iron homeostasis in SOD1^{G93A} ESMNs

C) Calcium regulators and lipogenesis genes identified by scTDA show common expression patterns in clusters on the kNN graph



Supplemental Figure 23. Motor neuron markers in the Day 3 and Day 7 populations of ESMNs. (expression log scale 0 -- Max)