



UNIVERSITAT DE
BARCELONA

Characterization of alternative polyadenylation sites at single cell resolution

Franz Ake



Aquesta tesi doctoral està subjecta a la llicència **Reconeixement 4.0. Espanya de Creative Commons.**

Esta tesis doctoral está sujeta a la licencia **Reconocimiento 4.0. España de Creative Commons.**

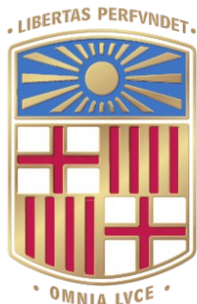
This doctoral thesis is licensed under the **Creative Commons Attribution 4.0. Spain License.**

Doctoral Thesis

Characterization of alternative polyadenylation at single-cell resolution

Author: Franz AKE

Supervisor: Dra Mireya PLASS



UNIVERSITAT_{DE}
BARCELONA



UNIVERSITAT DE
BARCELONA

FACULTAT DE MEDICINA I CIÈNCIES DE LA SALUT

PROGRAMA DE DOCTORAT EN BIOMEDICINA

Department of Genetics, Microbiology and Statistics - Faculty of
Biology.

2025

TESI DOCTORAL

**Characterization of alternative polyadenylation at
single-cell resolution**

*

**Caracterizació de la poliadenilació alternativa a
resolució de cèl·lula única**

Memòria presentada per

Franz AKE

Per optar al títol de

Doctor

Per la Universitat de Barcelona

Tesi doctoral realitzada sota la direcció de la Dra. Mireya Plass
i sota la tutoria de Dr Josep Francesc Abril Ferrando

Signat,

El Doctorand:

A handwritten signature in black ink, appearing to be "FAKE", written over a horizontal line.

Joy is the simplest form of gratitude.

Acknowledgments

Je tiens d'abord à exprimer toute ma gratitude à mon père, qui a indubitablement éveillé en moi le goût de la science, de la biologie, de la curiosité, et, surtout, l'importance d'entretenir un esprit libre qui remet sans cesse le monde en question pour mieux en faire sien. À ma douce mère, merci infiniment pour ton amour constant, ton soutien indéfectible et les innombrables sacrifices que tu n'as jamais hésité à effectuer pour tes garçons. À ma famille, Michael, Arthur, Edith, et tous mes cousins et cousines en Côte d'Ivoire et France, je vous remercie de votre présence dans ma vie qui constitue une joie et un réconfort. Je remercie aussi de tout cœur ma chère Tata Colette, qui est devenue avec le temps une seconde maman et une amie. Merci pour ta bienveillance et ton affection et tout ton soutien. Merci à Tonton Jacques, ainsi qu'à tous mes oncles de la région parisienne — Tonton David, Tonton Jacques Sika, Tonton Gervais, et bien d'autres — pour m'avoir accueilli comme un fils et constitué une seconde famille. Également, je voudrais adresser un mot particulier à Tata Ouraga et Tonton Bamba pour leur hospitalité, et leur aide et soutien à mon égard. À mon frère de cœur Yann, merci pour notre longue amitié qui a su traverser les années et toutes les péripéties chavilloises, ainsi qu'Ariane et toute ta famille. Je pense également à toute ma bande d'amis "la bande timal" en région parisienne, que je remercie pour les rires, et tous les souvenirs partagés. Enfin, À Jessica, Thierry et toute sa magnifique famille, je voudrais vous remercier pour votre amour, votre bienveillance et surtout tous les fous rires et incroyables moments passés ensemble.

Durant ce parcours universitaire, j'ai eu à traverser des périodes de doute profond, de démotivation totale, me sentant parfois illégitime dans le monde académique, comme un alien pas à sa place. Dans ces moments critiques, j'ai eu la chance de croiser le chemin du Professeur Costas Bouyioukos à mon université Paris Diderot, dont le regard bienveillant, le soutien et la confiance ont été déterminants. Merci de m'avoir tendu la main, d'avoir cru en moi et donné ma chance. Votre impact sur mon parcours dépasse largement ce que vous pouvez imaginer. Je tiens également à remercier chaleureusement le laboratoire PTER (CNRS) ainsi que le Dr Cosson pour leur accueil généreux et stimulant. Dans la même lignée, je tiens à remercier très affectueusement mon incroyable paire de superviseurs Dr Marie Binet et Dr Carole Guillaume, mes encadrantes pour mon stage final de Master au LNHE – EDF, pour leur incroyable disponibilité, leur bienveillance et leur soutien indéfectible, bien au-delà du cadre du stage. Vous avez accompagné mes premiers pas hésitants de bioinformaticien avec une humanité et une patience inoubliables. Merci à toute l'équipe de l'INRAe GET-IT à Toulouse pour votre accueil, votre professionnalisme et votre amitié. Mon passage parmi vous, bien que bref, m'a énormément apporté et m'a permis de renforcer mes compétences tout en m'ouvrant à de nouvelles perspectives.

Enfin, je tiens à adresser toute ma reconnaissance à la personne centrale de cette thèse : ma directrice, Mireya Plass. En arrivant à Barcelone, je t'ai dit que je voulais apprendre, progresser, devenir meilleur. Aujourd'hui, je peux dire sans hésiter que tu as joué un rôle fondamental dans cette transformation. Ton exigence scientifique, ton sens critique, ta rigueur et ta manière particulière de résoudre les problématiques qui se posent m'ont indubitablement marqué en

profondeur. Tu vas me manquer, autant que nos meeting hebdomadaires et longues soirées à IDIBELL passées à déboguer des erreurs de code que tu résolvais avec une élégance désarmante. Je souhaite également remercier tous mes collègues du laboratoire, sans qui ce travail n'aurait pas eu la même richesse ni la même humanité. Merci à Ana, pour ton soutien sur tous les plans, ta présence amicale et ta générosité. À Sandra, pour ta disponibilité et ton professionnalisme. À Marcel, pour ton aide technique précieuse et ta contribution directe à l'amélioration de SCALPEL. À Akshay, pour ton implication dans les analyses et ta collaboration. Et à tous les autres membres du labo que j'ai eu la joie de fréquenter : Nadia, Rafael, Natalie, Mohamed, Hussein, Sara ; merci pour vos échanges, vos sourires, vos encouragements.

Pour conclure, je remercie chaleureusement tous mes amis, collègues et collaborateurs du PCRMC et de l'IDIBELL, pour leur gentillesse, leur aide précieuse et leur présence tout au long de cette belle aventure humaine et scientifique.

Abstract

Alternative polyadenylation (APA) is a widespread mechanism of gene regulation that affects more than 70% of human genes. mRNA isoforms generated through APA have distinct 3' ends, which can affect mRNA regulation as well as the resulting proteins. APA is well known to be regulated during cell differentiation and it is a major source of gene regulation in the brain. Yet, it is not known till which extent APA contributes to the transcriptomic variability across individual cell populations. While bulk RNA sequencing has provided important insights into APA, it does not provide the resolution needed to resolve the cell-type-specific dynamics essential for understanding development and disease. The rapid development of 3' tagged-based single-cell RNA sequencing (scRNA-seq) has enabled the study of gene expression and the implementation of methods for describing isoform usage at single cell resolution. However, challenges such as dropout effects or technical biases arising from the experimental setup have to be surmounted. In this thesis, we developed a novel computational approach to characterize isoform diversity at single-cell resolution using 3'-tagged scRNA-seq data, thereby enhancing the understanding of APA dynamics at single-cell resolution. First, I systematically evaluated the impact of various single-cell transcriptomics sample preservation methods on neural and glial cells derived from induced pluripotent stem cells. Secondly, I implemented and benchmarked SCALPEL, a novel computational Nextflow workflow designed for quantifying isoforms using standard 3'-tagged scRNA-seq data. Our results show that while DMSO provides the highest cell quality in terms of RNA molecules and genes detected per cell, it strongly affects the cellular composition and induces the

expression of stress and apoptosis genes. In contrast, methanol fixed samples display a cellular composition similar to fresh samples and provide a good cell quality and little expression biases, indicating that methanol fixation is the method of choice for performing droplet-based single-cell transcriptomics experiments on neural cell populations. Furthermore, SCALPEL demonstrated high sensitivity and precision in quantifying isoforms, and the ability to detect differential isoform usage (DIU) in both synthetic and experimental datasets. In real datasets, SCALPEL predictions have a high agreement with other tools and can be experimentally validated. The use of SCALPEL on real datasets reveals novel cell populations undetectable using single-cell gene expression data, confirms known 3' UTR length changes during cell differentiation, and identifies cell-type specific miRNA signatures regulating isoform expression. Additionally, we show that SCALPEL improves isoform quantification using paired long- and short-read scRNA-seq data. Overall, SCALPEL expands the current scRNA-seq toolkit to explore post-transcriptional gene regulation across species, tissues, and technologies, advancing our understanding of gene regulatory mechanisms at the single-cell level.

Keywords:

Alternative polyadenylation

Single-cell RNA sequencing

Isoform quantification

Sample preservation

Computational method development

Resum

La poliadenilació alternativa (APA) és un mecanisme estès de regulació gènica que afecta més del 70% dels gens humans. Els isoformes d'ARNm generats mitjançant APA tenen extrems 3' diferents, fet que pot influir tant en la regulació de l'ARNm com en la proteïna resultant. L'APA és ben coneguda per estar regulada durant la diferenciació cel·lular i constitueix una font important de regulació gènica al cervell. Tot i això, encara no se sap fins a quin punt l'APA contribueix a la variabilitat transcriptòmica entre poblacions cel·lulars individuals. Encara que el RNA-seq a granel ha proporcionat informació rellevant sobre l'APA, no ofereix la resolució necessària per captar les dinàmiques específiques de cada tipus cel·lular, essencials per comprendre el desenvolupament i les malalties. El desenvolupament accelerat del RNA-seq de cèl·lula única basat en etiquetatge 3' ha permès estudiar l'expressió gènica i implementar mètodes per descriure l'ús d'isoformes a nivell de cèl·lula individual. No obstant això, cal superar desafiaments com els efectes de dropout o els biaixos tècnics derivats del disseny experimental. En aquesta tesi, hem desenvolupat un nou enfocament computacional per caracteritzar la diversitat d'isoformes a resolució de cèl·lula única utilitzant dades de RNA-seq de cèl·lula única etiquetades al 3', amb l'objectiu de millorar la comprensió de les dinàmiques d'APA a aquest nivell. En primer lloc, vaig avaluar sistemàticament l'impacte de diversos mètodes de preservació de mostres en transcriptòmica de cèl·lula única aplicats a cèl·lules neurals i glials derivades de cèl·lules mare pluripotents induïdes (iPSC). En segon lloc, vaig implementar i avaluar SCALPEL, un nou flux de treball en Nextflow per a la quantificació d'isoformes mitjançant dades convencionals de

scRNA-seq etiquetades al 3'. Els nostres resultats mostren que, tot i que el DMSO proporciona una qualitat cel·lular elevada en termes de molècules d'ARN i gens detectats per cèl·lula, afecta fortament la composició cel·lular i indueix l'expressió de gens relacionats amb estrès i apoptosi. En canvi, les mostres fixades amb metanol mostren una composició cel·lular similar a les mostres fresques i presenten una bona qualitat transcriptòmica i pocs biaixos d'expressió, indicant que la fixació amb metanol és el mètode preferible per a experiments de scRNA-seq basats en gotes per a poblacions cel·lulars neurals. A més, SCALPEL ha demostrat una gran sensibilitat i precisió en la quantificació d'isoformes i en la detecció d'ús diferencial d'isoformes (DIU) tant en conjunts de dades sintètics com experimentals. En dades reals, les prediccions de SCALPEL tenen una alta concordança amb altres eines i poden ser validades experimentalment. L'ús de SCALPEL en dades reals revela poblacions cel·lulars noves no detectables mitjançant dades d'expressió gènica de cèl·lula única, confirma canvis coneguts en la longitud dels 3' UTR durant la diferenciació cel·lular i identifica signatures específiques de miARN que regulen l'expressió d'isoformes segons el tipus cel·lular. Addicionalment, demostrem que SCALPEL millora la quantificació d'isoformes utilitzant dades aparellades de seqüenciació curta i llarga de cèl·lula única. En conjunt, SCALPEL amplia les capacitats actuals de la transcriptòmica de cèl·lula única per explorar la regulació gènica post-transcripcional en diferents espècies, teixits i tecnologies, i fa avançar la nostra comprensió dels mecanismes reguladors a nivell de cèl·lula individual.

Paraules clau:

Poliadenilació alternativa

Seqüenciació d'ARN de cèl·lula única

Quantificació d'isoformes

Preservació de mostres

Desenvolupament de mètodes computacionals

Table of Contents

Acknowledgments	5
Abstract	9
Resum	11
Table of Contents	15
List of figures	23
List of tables	27
List of abbreviations	31
1. Introduction	35
1.1. Introduction to alternative polyadenylation (APA).....	37
1.1.1. APA mechanism	37
1.1.2. Classes of APA	37
1.1.3. Regulation of APA.....	39
1.1.4. Implications of APA in biological processes	45
1.1.5. Implications of APA in diseases	48
1.1.6. Characterization of APA: from low to high-throughput approaches.....	49
1.2. Study of gene expression using single-cell transcriptomics	53
1.2.1. Evolution of scRNA-seq experiments	53
1.2.2. Basic scRNA-seq workflow.....	58

1.2.3.	Library Construction Strategies	61
1.2.4.	Sequencing methods.....	62
1.3.	Challenges in the study of gene expression in single-cell data	64
1.3.1.	Dropout and batch effects	64
1.3.2.	Impact of cell dissociation on single-cell data.....	65
1.3.3.	Impact of sample preservation methods on the gene expression	66
1.4.	Characterization of APA at single-cell resolution	69
1.4.1.	Emergence of methods for APA profiling from single-cell data.....	69
1.4.2.	Computational methods for APA characterization	69
2.	Objectives	75
2.1.	Objective 1: Optimize single-cell transcriptomics protocols to work with iPSC-derived neuron and glial cells	77
2.2.	Objective 2: Implementation of a new computational method for the quantification of 3' end isoforms using scRNA-seq datasets.....	77
3.	Material and Methods.....	79
3.1.	Objective 1: Optimize single-cell transcriptomics protocols to work with iPSC-derived neuron and glial cells	81
3.1.1.	Sample preservation methods.....	81
3.1.2.	Single-cell RNA-seq library preparation	82
3.1.3.	scRNA-seq data pre-processing.....	83

3.1.4.	Filtering of low-quality cells and doublets	84
3.1.5.	Identification of cell populations.....	84
3.1.6.	Stress and apoptosis gene signatures	85
3.1.7.	Cell composition analysis	86
3.1.8.	Differential expression analysis	86
3.1.9.	Correlation analysis	87
3.2.	Objective 2: Implementation of a new computational method for the quantification of 3' end isoforms using scRNA-seq datasets.....	89
3.2.1.	Annotation preprocessing.....	89
3.2.2.	Read preprocessing	89
3.2.3.	Quantification of isoforms at the single-cell level.....	91
3.2.4.	Isoform entropy and gene information content calculation	93
3.2.5.	Detection of differential isoform usage between cell clusters	96
3.2.6.	Average 3' UTR length calculation	96
3.2.7.	Generation of read coverage plots	97
3.2.8.	Generation of synthetic data for benchmark analysis.....	97
3.2.9.	Analysis of mouse spermatogenesis 10x scRNA-seq data	99
3.2.10.	Differentiation and characterization of iPSCs and NPCs using Drop-seq technology.....	100

3.2.11.	scRNA-seq analysis of Drop-seq data.....	101
3.2.12.	Benchmark analysis	102
3.2.13.	Isoform validation using nested PCRs.....	103
3.2.14.	Identification of miRNA signatures in differentially regulated isoforms	104
3.2.15.	Identification of novel cell populations in the mouse 10x dataset.....	104
3.2.16.	Differential gene expression analysis between RS populations	105
3.2.17.	GO term enrichment analysis	106
3.2.18.	Trajectory inference analysis	106
3.2.19.	Analysis of paired short and long scRNA-seq datasets.	106
3.2.20.	Statistical analyses	108
3.2.21.	Data availability	108
3.2.22.	Code availability	109
4.	Results	111
4.1.	Objective 1: Optimize single-cell transcriptomics protocols to work with iPSC-derived neuron and glial cells	113
4.1.1.	Impact of sample preservation methods on single-cell library quality and cell composition.....	113
4.1.2.	Preservation methods induce biases in the gene expression profile of cells	121

4.2. Objective 2: Implementation of a computational method for characterization of APA with scRNA-seq datasets	127
4.2.1. SCALPEL: A pipeline for isoform quantification at single- cell resolution.....	127
4.2.2. Benchmarking of SCALPEL	129
4.2.3. Isoform-based clustering enables the identify a novel cell population	136
4.2.4. Investigation of APA profiles and 3' end isoform variability across the cell types	140
4.2.5. SCALPEL improves isoform quantification of novel isoforms predicted using single-cell long read sequencing	145
5. Discussion	149
5.1. Objective 1: Optimize single-cell transcriptomics protocols to work with iPSC-derived neuron and glial cells	152
5.1.1. Sample preservation methods impact the library quality in single-cell experiments	152
5.1.2. DMSO cryopreservation alters cell-type composition....	153
5.1.3. Preservation methods induce stress and apoptotic signatures	154
5.2. Objective 2: Implementation of a computational method for characterization of APA with scRNA-seq datasets	157

5.2.1.	SCALPEL, a computational framework for isoform quantification using single-cell data.....	157
5.2.2.	SCALPEL accurately quantifies isoform and outperform published tools for APA characterization in single-cell data	157
5.2.3.	SCALPEL isoform quantification conserves cell-type specificity and enable to recover novel cell types.....	158
5.2.4.	SCALPEL identifies global APA dynamic and isoform usage change across cell-types	159
6.	Conclusions.....	163
7.	References	167
8.	Appendices	193
8.1.	Supplementary Figures	195
8.2.	Supplementary Tables	211
8.3.	Publications	228

List of figures

Figure 1. Categories of APA.

Figure 2. Overview of RNA 3' end processing factors.

Figure 3. Consequences of APA regulation.

Figure 4. Functional Consequences of tandem 3' UTR-APA.

Figure 5. Evolution in the scaling of scRNA-seq experiments.

Figure 6. 10x single-cell encapsulation

Figure 7. Main steps in single-cell transcriptomics.

Figure 8. Overview of main scRNA-seq library preparation approaches.

Figure 9. Illustration of the peak calling based methods for PAS identification using 3'enriched scRNA-seq data.

Figure 10. Illustration of the distribution based methods for PAS identification using 3'enriched scRNA-seq data.

Figure 11. Systematic comparison of the effects of cellular preservation on library quality.

Figure 12. scRNA-seq identifies diverse progenitor and immature neuron populations.

Figure 13. DMSO cryopreservation impacts the cell composition of scRNA-seq datasets.

Figure 14. Computational analysis identifies fixation-related biases in cell clustering.

Figure 15. Fixation/preservation methods induce sample-specific expression biases.

Figure 16. SCALPEL pipeline quantifies transcript isoforms at the single-cell level.

Figure 17. Benchmark of SCALPEL using synthetic data.

Figure 18. Number of correctly identified features.

Figure 19. SCALPEL predictions on iPSC and NPC data can be experimentally validated.

Figure 20. Isoform quantification recapitulates 3' UTR shortening during mouse spermatogenesis.

Figure 21. SCALPEL predictions on iPSC and NPC reflect miRNA function.

Figure 22. SCALPEL provides isoform quantification on individual cells using paired short and long scRNA-seq datasets.

Supplementary figure 1. Distance distribution of reads to annotated 3' ends of genes with only one expressed isoform.

Supplementary figure 2. Detailed SCALPEL workflow.

Supplementary figure 3. Differences between gene and isoform-based clustering solutions on the mouse spermatogenesis dataset.

Supplementary figure 4. Cell preservation induces biases in gene expression that affect sample comparison.

Supplementary figure 5. Comparison of the capture efficiency of the 2 bead batches used in the project.

Supplementary figure 6. Cell composition biases across samples.

Supplementary figure 7. Additional feature plots validating the identity of the identified clusters.

Supplementary figure 8. Gene expression correlation across samples.

Supplementary figure 9. Cell-type expression correlation highlights fixation biases in cell clustering.

Supplementary figure 10. Additional plots on the benchmarking analysis.

Supplementary figure 11. Benchmarking of APA and isoform quantification tools on a 10x dataset of mouse spermatogenesis.

Supplementary figure 12. Benchmarking of APA and isoform quantification tools on a Drop-seq dataset containing iPSCs and NPCs.

Supplementary figure 13. Experimental validation of isoform changes between iPSCs and NPCs.

Supplementary figure 14. Additional information on the isoform diversity on the mouse spermatogenesis 10x dataset.

Supplementary figure 15. Additional details on the scRNA-seq 10x datasets of mouse brain with paired long and short read data.

List of tables

Table 1. Cell quality statistics

Table 2. Summary of the effects of preservation methods on single-cell libraries

Supplementary table 1. Cumulative percentage of cells discarded using each of the different parameters used to remove low quality cells

Supplementary table 2. Parameters and number of doublets identified using DoubletFinder

Supplementary table 3. Differentially expressed genes per cluster for each fixation/preservation method

Supplementary table 4. Types of genes simulated

Supplementary table 5. Characteristics of tools included in the benchmarking

Supplementary table 6. Top 20 gene DGE markers for 10x scRNA-seq data from mouse hippocampus and visual cortex

Supplementary table 7. Top20 isoform iDGE markers for 10x scRNA-seq data from mouse hippocampus and visual cortex

Supplementary table 8. DIU genes identified by each benchmarked tool

Supplementary table 9. Execution performance on the simulated scRNA-seq dataset

Supplementary table 10. Details on the number of predictions and technical details on the mouse spermatogenesis 10x dataset

Supplementary table 11. DIU genes between iPSCs and NPCs

Supplementary table 12. Gene markers on the high-resolution clustering analysis of mouse spermatogenesis dataset

Supplementary table 13. Isoform markers on the high-resolution clustering analysis of mouse spermatogenesis dataset

Supplementary table 14. GO-term enrichment analysis in RS6 vs other RS cells

Supplementary table 15. DIU genes between RS6 population and other RS populations

Supplementary table 16. DIU genes identified in the mouse spermatogenesis dataset

Supplementary table 17. NPC miRNAs evaluated as isoform expression regulators in NPCs

Supplementary table 18. Stats of detected feature from 10x short read and PacBio long read data

Supplementary table 19. Top 20 gene DGE markers for 10x scRNA-seq data from mouse hippocampus and visual cortex

Supplementary table 20. Top20 isoform iDGE markers for 10x scRNA-seq data from mouse hippocampus and visual cortex

Supplementary table 21. Differential isoform usage (DIU) genes between the different cell-types using SCALPEL in 10x scRNA-seq data from Joglekar et al.

List of abbreviations

APA: Alternative Polyadenylation

UTRs: Untranslated Regions

PAS: Polyadenylation Site

ATE: Alternative Terminal Exon

USEs: Upstream Sequence Elements

DSEs: Downstream Sequence Elements

CTD: Carboxy-Terminal Domain

RNA Pol II: RNA Polymerase II

PAP: Poly(A) Polymerase

cUTR: Constitutive UTR

aUTR: Alternative UTR

RBPs: RNA-Binding Proteins

miRNA: MicroRNA

lncRNAs: Long Non-Coding RNAs

CPSF: Cleavage and Polyadenylation Specificity Factor

CstF: Cleavage Stimulatory Factor

CFIm: Cleavage Factors I and II

AREs: AU-Rich Elements

TTP: Tristetraprolin

IPEX: Immunodysregulation Polyendocrinopathy Enteropathy X-Linked

ALS: Amyotrophic Lateral Sclerosis

3' RACE: 3' Rapid Amplification of cDNA Ends

EST: Expressed Sequence Tag

NGS: Next-Generation Sequencing

RNA-Seq: RNA Sequencing

scRNA-seq: Single-Cell RNA-Seq

cDNA: Complementary DNA

STRT-seq: Single-Cell Tagged Reverse Transcription Sequencing

TSO: Template-Switching Oligos

MARS-seq: Massive Analysis of RNA-Seq

STRT-seq-2i: A Specific Protocol Variant of STRT-seq

GEMs: Gel Beads-in-Emulsion

RT: Reverse Transcription

FACS: Fluorescence-Activated Cell Sorting

UMIs: Unique Molecular Identifiers

PCR: Polymerase Chain Reaction

IVT: In Vitro Transcription

DGE: Digital Gene Expression

PCA: Principal Component Analysis

ACME: Acetic-Methanol

PFA: Paraformaldehyde

DSP: Dithio-Bis-Succinimidyl Propionate

DPBS: Dulbecco's Phosphate Buffered Saline

DMSO: Dimethyl Sulfoxide

DTT: Dithiothreitol

PCs: Principal Components

UMAP: Uniform Manifold Approximation and Projection

DEGs: Differentially Expressed Genes

BED: BED File Format

IP: Internal Priming

iDGE: Isoform Digital Gene Expression Matrix

DIU: Differential Isoform Usage

GTF: GTF File Format

FASTA: FASTA File Format

FASTQ: FASTQ File Format

GEO: Gene Expression Omnibus

SC: Spermatocytes

RS: Round Spermatids

ES: Elongated Spermatids

iPSCs: Induced Pluripotent Stem Cells

NPCs: Neural Progenitor Cells

GO: Gene Ontology

NMD: Nonsense-Mediated mRNA Decay

1. Introduction

1.1. Introduction to alternative polyadenylation (APA)

1.1.1. APA mechanism

Alternative polyadenylation (APA) is a widespread post-transcriptional mechanism that enables the generation of multiple mature RNA isoforms with distinct 3' ends, usually affecting the 3' untranslated regions (UTRs) of a single gene¹. This process occurs during the RNA molecule maturation following DNA transcription and involves the cleavage and the addition of a poly(A) tail at a defined polyadenylation site (PAS). PAS are predominantly located within the 3' UTR region and are classified as proximal or distal based on their relative distance from the STOP codon². This regulated process allows cells to produce mRNA isoforms with different 3' UTRs that can be differentially regulated, thereby affecting stability, nuclear export, and translation³. While APA affects coding and many types of non-coding RNAs⁴, its functional consequences have been mostly studied on mRNAs. Thus, I will focus on the role of APA on RNA throughout this thesis.

1.1.2. Classes of APA

APA generates distinct mRNA isoforms by selecting different PAS within a gene, resulting in variations in their 3' end regions. Depending on the location of selected PAS, APA can be classified into four general classes, each with unique regulatory implications. 3' UTR tandem APA is the most prevalent form, occurring in approximately 70% of mammalian genes⁵. In this type of APA, multiple PAS are located within the same terminal exon of the pre-mRNA. The selection of proximal or distal sites generates mRNA isoforms with varying 3' UTR lengths

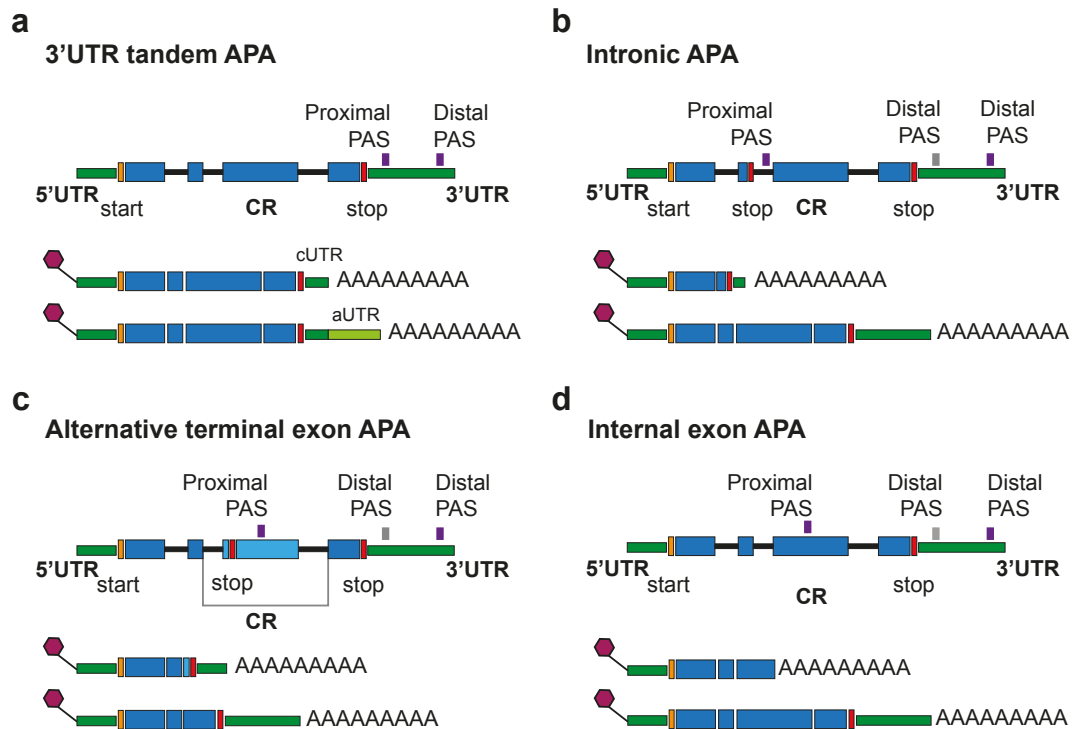


Figure 1. Categories of APA a. Tandem 3' UTR-APA containing two or more poly(A) sites in the 3' untranslated region b. Intronic APA occurring in the introns c. Splicing APA (alternative terminal exon APA) possessing a proximal PAS in the last exon and resulting in internal exon skipping d. Internal exon APA generating a 3' UTR lacking isoform via PAS usage in the upstream exon (adapted from Zhang Y. et al.¹⁴)

(Figure 1a). Shorter 3' UTRs, produced by proximal PAS usage, may lack regulatory elements such as microRNA (miRNA) and RBP binding sites, potentially increasing mRNA stability and translation efficiency. In contrast, longer 3' UTRs, resulting from distal PAS usage, can include numerous regulatory elements that modulate mRNA decay, localization, and translation^{1,3}. Intronic APA events occur when the cleavage takes place at a PAS located within intronic regions, resulting in truncated mRNA isoforms. These truncated isoforms encode shorter protein variants or non-coding RNAs, contributing to the regulation of gene expression or protein function (Figure 1b). Alternative terminal exon (ATE)

APA events involve the selection of different terminal exons, each harboring distinct PAS. This type of APA event not only alters the 3' UTR but can also affect the coding region of the mRNA, leading to the production of diverse protein isoforms with different functions and/or localizations (Figure 1c)^{6,7}. Finally, internal exon APA events involve the selection of PAS within internal exons. Similar to ATE APA and intronic APA events, this type of APA can alter the coding region of the mRNA leading to the generation of protein variants with distinct functional domains or regulatory properties (Figure 1d)^{8,9}.

1.1.3. Regulation of APA

APA is regulated through a complex interplay of cis-regulatory elements embedded within the RNA sequence that determine PAS selection, and trans-acting protein regulators that mediate or modulate the RNA molecule processing. These regulatory layers coordinate the recognition and selection of PAS, ultimately influencing isoform diversity^{1,7}.

1.1.3.1. Cis regulatory elements of APA

APA regulation relies on cis-regulatory elements embedded within the pre-mRNA sequence, near PAS, which serve as binding platforms for the polyadenylation machinery. The core cis-regulatory element is the canonical PAS signal, typically the hexanucleotide sequence *AAUAAA* located approximately 10–30 nucleotides upstream of the cleavage site (Figure 2a). This sequence is the most conserved and efficient motif guiding the core polyadenylation machinery complex for cleavage and polyadenylation of the immature RNA molecule. Variants such as *AUUAAA* or *UAUAAA* can be recognized by the polyadenylation machinery but

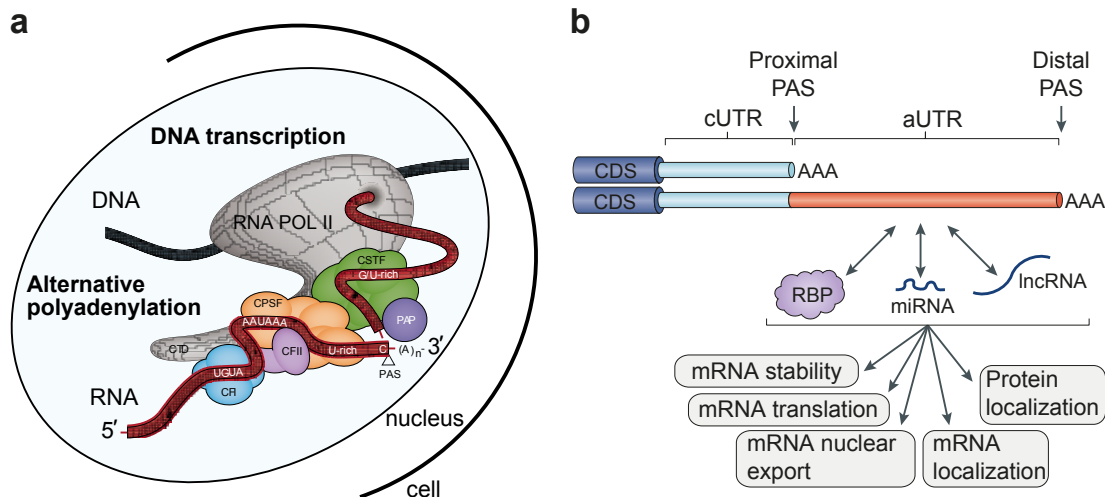


Figure 2. Overview of RNA 3' end processing factors a. The 3' end processing complex is composed mainly of four subcomplexes CPSF, CFI, CFII and CSTF, which interact with the carboxy-terminal domain (CTD) of RNA Polymerase II (RNA Pol II). Subcomplexes interact with the indicated sequence motifs ('AAUAAA' being the canonical poly(A) signal) in nascent RNAs to dynamically determine the site of cleavage and polyadenylation, located preferentially after a cytosine (indicated by 'C'). A poly(A) polymerase (PAP) synthesizes the poly(A) tail. b. Alternative polyadenylation (APA) generates two mRNA isoforms with different 3' UTRs, referred to as 3' UTR-APA. The region upstream of the proximal polyadenylation site (PAS) is shared by both the short (top) and long (bottom) isoforms, called the constitutive UTR (cUTR), while the downstream region is present only in the long isoform, known as the alternative UTR (aUTR). Interactions between the aUTR and RNA-binding proteins (RBPs), microRNAs (miRNAs), or lncRNAs can lead to diverse functional outcomes (adapted from Gruber & Zavolan ^{1,3}).

generally exhibit reduced cleavage efficiency and require additional *cis*-regulatory elements or cooperating trans-acting factors to ensure functional PAS activation and expression¹⁰.

The surrounding upstream sequence elements (USEs) and downstream sequence elements (DSEs) also modulate PAS recognition¹¹ (Figure 2a). DSEs

are typically GU or U-rich motifs that contribute to cleavage site positioning, while USEs may enhance complex assembly and processing efficiency³. The efficiency of the cleavage and polyadenylation at a given PAS is strongly influenced by the binding affinity of the elements of the core polyadenylation machinery complex with these cis-regulatory elements. For instance, a canonical AAUAAA motif in combination with a GU-rich DSE ensures high cleavage efficiency and robust PAS usage, whereas suboptimal configurations, such as PAS variants or weakly defined flanking sequences, reduce the binding affinity of core processing factors and result in inefficient processing¹². Collectively, the strength and spatial arrangement of cis-regulatory elements govern the efficacy of the cleavage and polyadenylation reaction, thereby influencing alternative PAS selection and shaping APA isoform diversity¹³.

1.1.3.2. Trans-acting regulators of APA

APA is tightly regulated by trans-acting factors that modulate the recognition and selection of PAS. A key layer of this regulation is constituted by the core cleavage and polyadenylation machinery, which consists of several multi-protein complexes (Figure 2a). They are constituted mainly of the cleavage and polyadenylation specificity factor (CPSF), which recognizes the canonical AAUAAA motif within the pre-mRNA molecule and catalyzes its cleavage. The cleavage stimulatory factor (CstF), which binds to GU/U-rich sequences downstream of the PAS and works synergistically with CPSF to position the cleavage site, and the additional cleavage factors I and II (CFIm and CFII) to stabilize the polyadenylation machinery complex and modulate PAS selection, particularly CFIm, which enhances usage of distal PAS through its interaction with

UGUA motifs. Additional proteins, such as Symplekin acts as a scaffold to facilitate the assembly of the machinery complex, while poly(A) polymerase (PAP) catalyzes the poly(A) tail addition. The concentration and composition of these factors influence the choice of the PAS and broadly shape the diversity of APA isoforms^{1,14}. Changes in the abundance of these proteins, particularly CPSF subunits (e.g., CPSF160, CPSF73), CstF64, and CFIm25 have been shown to shift PAS usage¹⁵. For example, depletion of CFIm25 promotes usage of proximal PAS, leading to widespread shortening of 3' UTRs, whereas its overexpression favors a distal PAS usage and generation of longer 3' UTR isoforms^{16,17}.

Beyond the core poly(A) machinery, RNA binding proteins (RBPs) play a crucial role in modulating PAS by interacting with specific cis-regulatory elements within the pre-mRNA near PAS, thus influencing their accessibility and recruitment by poly(A) machinery proteins. These interactions can either stabilize or repress the usage of certain PAS leading to the generation of alternative mRNA isoforms with distinct 3' UTR regions, impacting their stability, localization, and translation efficiency in proliferative, stress-responsive or cell differentiation context^{18,19}. For instance, under cellular stress or immune activation, RBP HuR (ELAV1) modulates APA in T cells, neurons, and tumor cells by stabilizing transcripts with longer 3' UTRs. This regulation contributes to buffering against stress-induced mRNA decay and sustains expression of genes involved in survival and immune response²⁰. Similarly, RBP PTBP1 plays a key role in cell reprogramming or differentiation by promoting proximal PAS usage, leading to widespread 3' UTR shortening. PTBP1 has been shown to mediate this effect in B cells, neuronal progenitors, and glioma cells, where its upregulation correlates with transcriptome remodeling characteristic of a more proliferative or dedifferentiated

state^{21,22}. Furthermore, the elongation rate of RNA polymerase II and the chromatin state surrounding the PAS can significantly affect APA. Slow transcription elongation rates can favor the usage of proximal PAS, while rapid elongation rates and open chromatin structures are associated with distal PAS usage⁶.

1.1.3.3. Molecular function of APA

APA alters the composition of the 3' UTR, which contains binding sites for RBPs, microRNAs (miRNAs), and long non-coding RNAs (lncRNAs) that are key regulators of mRNA fate. These interactions modulate transcript stability, localization, and translation^{6,7,23} (Figure 2b). For instance, the RBP HuR binds to AU-rich elements (AREs) within the 3' UTR, enhancing mRNA stability by protecting it from degradation pathways²⁴. In contrast, Tristetraprolin (TTP) recognizes similar AREs but promotes mRNA decay by recruiting deadenylase complexes, which shorten the poly(A) tail, leading to rapid degradation of the mRNA²⁵.

Additionally, APA can affect the translation efficiency of mRNA isoforms and the subcellular localization of the encoded proteins^{20,26}. A well-characterized example is the CD47 gene, which generates isoforms with distinct 3' UTRs. The long 3' UTR isoform of this gene contains HuR-binding sites required for efficient translocation of the CD47 protein from the endoplasmic reticulum to the plasma membrane (Figure 3a). This process is mediated by the recruitment of the chromatin-associated factor SET and activation of RAC1 at the site of translation.

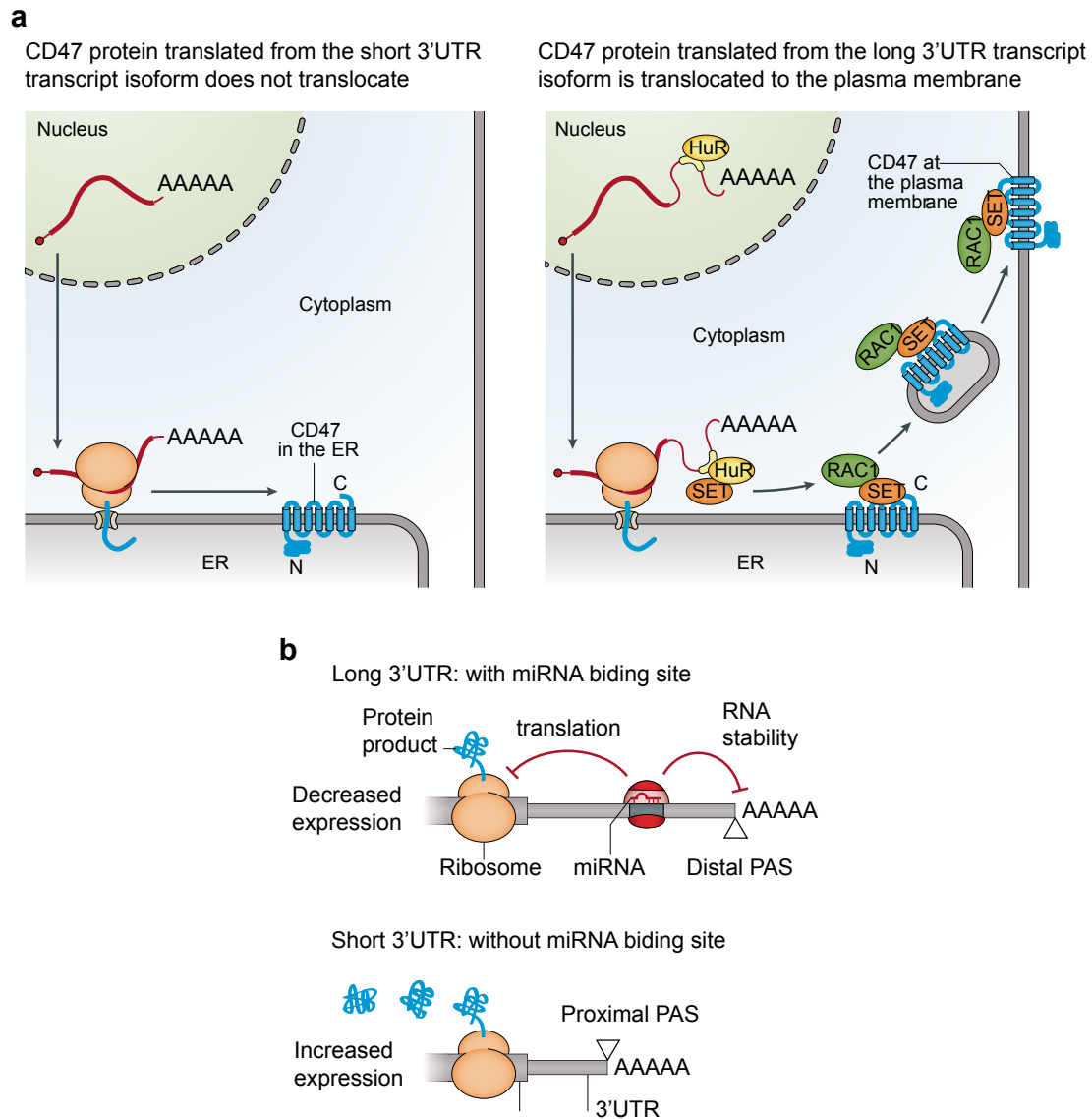


Figure 3. Consequences of APA regulation. **a.** Cis-regulatory elements in 3' UTRs can mediate the cellular localization of the encoded protein⁶⁴. HuR binding sites in the 3' UTR of the long isoform of CD47 are necessary and sufficient for the translocation of CD47 protein from the endoplasmic reticulum (ER) to the plasma membrane. The long 3' UTR isoform recruits a protein complex which contains HuR and SET to the site of translation, facilitating the joining of active RAC1 and the translocation of the CD47 transmembrane protein to the plasma membrane. Thus, the long 3' UTR of the CD47 isoform acts as a scaffold for a protein complex that mediates the translocation of CD47 from the ER to the plasma membrane **b.** Long 3' UTRs contain more regulatory elements that influence the stability and translation rate of the isoforms than short 3' UTRs. The presence of a microRNA (miRNA) binding site in a long 3' UTR can destabilize the mRNA and decrease the expression of the protein (adapted from Gruber et al ³).

In contrast, the short isoform lacks these regulatory elements, resulting in intracellular retention of the CD47 protein and reduced surface expression (Figure 3a). Since membrane localization is necessary for CD47 to function as an anti-phagocytic signal, APA-driven isoform selection directly modulates its immune regulatory role²⁰.

Moreover, miRNAs modulate the mRNA fate by binding to complementary sequences within the 3' UTR, leading to either translational repression or mRNA degradation (Figure 3b)²⁷. The availability of miRNA binding sites in APA isoforms influences their fate. Longer 3' UTRs generally harbor more miRNA binding sites, while shorter isoforms can evade miRNA-mediated repression leading to increased translation and higher protein output²⁸.

1.1.4. Implications of APA in biological processes

1.1.4.1. Tissue and Cell-Type Specificity of APA

APA exhibits widespread tissue specificity, reflecting a diversity in the profile of isoform usage across organs. Multiple studies have demonstrated that APA patterns vary across tissues, with distinctive PAS selection across diverse organs resulting in characteristic 3'UTR lengths^{29–31}. For instance, mRNA isoforms in the nervous system preferentially use distal PAS, generating isoforms with longer 3'UTRs, whereas in the placenta, ovaries and blood, they show an increased usage of proximal PAS³². In cardiac tissues, particularly under stress, there is a shift towards proximal PAS usage generating isoforms with shorter 3' UTRs that enhance their stability and translation of stress-response proteins³³. Comparative analysis of the tissues of five different mammals indicated that APA profiles in

different tissues are well conserved across species^{34–36}. Importantly, ubiquitously expressed genes are more likely to express APA isoforms in different tissues than genes with a restricted tissue expression, raising the possibility that 3' UTR tandem APA has an important role in tissue-specific regulation of these genes³⁷. Moreover, recent studies using single-cell transcriptomics have revealed that APA variation is not only tissue-specific but can also distinguish cell states/types within the same tissue^{38,39}. For example, in the developing nervous system, neuronal progenitors and mature neurons exhibit distinct PAS usage patterns, reflecting a developmental transition in post-transcriptional control mechanisms^{32,37}.

1.1.4.2. APA dynamics in developmental and proliferative processes

APA affects a wide range of biological processes in eukaryotes, including cellular differentiation, proliferation and stress response^{40–42}. Multiple studies have shown that APA is globally regulated according to cellular proliferation and differentiation status of cells (Figure 4). In mammals, proliferating cells show a preferential usage of proximal PAS, resulting in shorter 3' UTRs that lack binding sites for miRNAs or RBPs, and can evade repression, thus promoting cell growth^{1,3}. This APA regulation pattern is observed in cancer, stem cells, and reprogrammed somatic cells^{1,3}.

In contrast, mRNAs with longer 3' UTRs, which contain more regulatory sites, are associated with differentiated and highly specialized cell types, such as neurons^{32,42,43}. During embryonic development, for instance, studies in mammals, zebrafish, and *Drosophila melanogaster* have demonstrated a

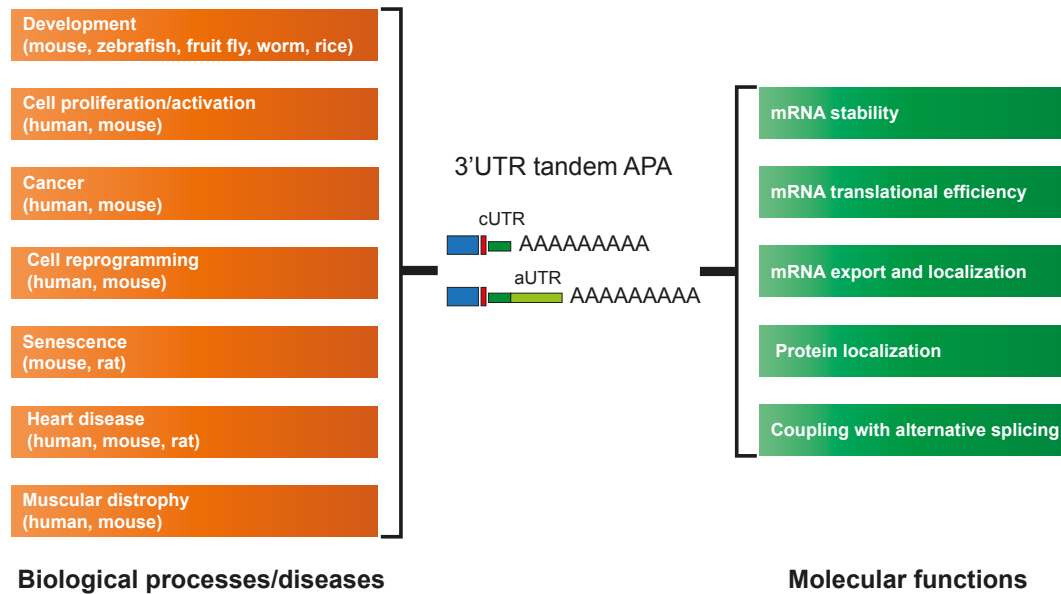


Figure 4. Functional Consequences of tandem 3' UTR-APA. This figure illustrates the physiological/pathological processes (left), and molecular functions (right) influenced by UTR-APA. APA plays a role in various biological processes, including development, cell proliferation, activation, reprogramming, and senescence. Additionally, APA is implicated in several pathological conditions, such as cancer, heart diseases, muscular dystrophy, and amyotrophic lateral sclerosis. On the molecular level, APA affects gene expression by influencing mRNA stability, translation efficiency, export, localization, and protein localization, and it interacts with alternative splicing (adapted from Chen et al¹³).

progressive 3' UTR lengthening as cells transition from a proliferative to a differentiated state^{39,44}. Largely, APA reprogramming towards longer isoforms has been described in cell cycle exit and acquisition of differentiated cell identity^{43,45}. Interestingly, the opposite was observed during spermatogenesis. As male germ cells differentiate, particularly in post-meiotic spermatids, there is a shift toward proximal PAS usage leading to widespread 3'UTR shortening^{46,47}. These observations highlight the regulatory importance of APA in gene expression

programs during development, cellular differentiation, and physiological adaptation.

1.1.5. Implications of APA in diseases

PAS alterations have been implicated in a wide range of human diseases, including cancers, neurodegenerative disorders, hematological conditions, and autoimmune syndromes^{48,49}. In cancer, recent studies show a preferential usage of proximal PAS, resulting in mRNA isoforms with shortened 3' UTRs that evade miRNA-mediated repression and enhance oncogene expression, promoting tumor growth and metastasis^{50–52}. These alterations, observed in a broad range of cancers, are often driven by changes in the expression of the core poly(A) machinery factors such as CSTF2⁵³ and CFIm25¹⁶, leading to a broad change in isoform usage. Additionally, intronic PAS activation can yield truncated protein isoforms that lack tumor suppressor functions or acquire oncogenic properties, as seen in chronic lymphocytic leukemia, where tumor suppressor proteins such as *Dicer* and *FOXN3* are affected⁵¹.

Moreover, mutations in PAS can directly impair 3' end processing, contributing to disease phenotypes as described in hematological disorders⁵⁴. For instance, a point mutation in the canonical PAS of the *FOXP3* gene (AATAAA to AATGAA) leads to reduced transcript stability and is causative in immunodysregulation polyendocrinopathy enteropathy X-linked (IPEX) syndrome⁵⁴. In neurodegenerative diseases such as Parkinson's disease or Amyotrophic lateral sclerosis (ALS), several studies indicated that APA dysregulation alters transcript isoform expression, contributing to disrupted neuronal homeostasis^{55,56}. For instance, Rhinn et al, reported that specific RNA transcript isoforms of α -synuclein

with an extended 3' UTRs, appear selectively linked to pathological processes, relative to shorter α -synuclein transcripts, impacting the accumulation of the resulting protein and changes in its subcellular localization from synaptic terminals to mitochondria⁵⁷. Although the specific mechanisms differ by disease, common features include isoform-specific changes in mRNA stability, localization, and protein output, highlighting APA as a critical post-transcriptional regulatory layer with pathological consequences.

1.1.6. Characterization of APA: from low to high-throughput approaches

1.1.6.1. Early approaches: From gene-specific to early global APA profiling

The study of APA has evolved significantly over the past two decades, from low-throughput gene-specific analyses to large-scale transcriptome-wide profiling. Early investigations relied on low-throughput molecular techniques, including Northern blotting, RNase protection assays, Sanger sequencing of cDNA clones, and 3' rapid amplification of cDNA ends (3' RACE)^{58–60}. These methods enabled the detection of mRNA isoforms that differ in their 3' UTRs or terminal exons due to the use of alternative PAS. Notably, such approaches were instrumental in demonstrating that multiple mRNA isoforms could arise from a single gene locus.

However, these methods were labor-intensive, low-resolution, and unsuitable for large-scale analysis. A major shift occurred with the advent of expressed sequence tag (EST) databases and microarray technologies, which provided the first global perspectives on APA usage in mammals. By aligning thousands of

EST sequences to genomic references, researchers were able to infer the locations of PAS across diverse tissues and species. Similarly, microarray platforms were adapted to measure APA isoforms by designing probes that target different regions of the 3' UTR. Nonetheless, both EST and microarray-based approaches were affected by significant technical limitations. EST data were influenced by cloning biases and lacked full-length coverage, limiting the resolution and quantifiability of APA isoform usage. Similarly, microarrays suffered from probe design constraints, incomplete isoform coverage, and limited dynamic range, restricting the detection of low-abundance isoforms or distinguishing closely spaced PAS^{5,61}.

1.1.6.2. The NGS Revolution: High-throughput bulk APA profiling

1.1.6.2.1. Standard RNA Sequencing

The emergence of next-generation sequencing (NGS) marked a turning point, enabling the transition from low to high-throughput APA profiling. These technologies allowed profiling millions of cells from a single sample or tissue⁶². In comparison to microarrays, RNA sequencing (RNA-Seq) provides a comprehensive transcriptome view of the biological sample without prior gene annotation^{63,64}. Although standard RNA-seq protocols were not initially developed to capture PAS usage, they provided sufficient resolution to infer APA events retrospectively. Standard RNA-seq datasets have been repurposed to infer changes in PAS usage by analyzing read coverage patterns across 3' UTRs. This approach led to the development of computational framework for characterizing APA from bulk RNA-seq data, with tools such as DaPars⁵³, APATrap⁶⁵, or QAPA⁶⁶.

These methods allow to estimate APA changes from differential read density between proximal and distal 3' UTR regions.

1.1.6.2.2. Specialized 3' end enriched RNA-seq methods

To improve specificity and resolution of PAS mapping, several methods were developed to selectively enrich the read coverage at the 3' end region of the isoforms. Several specialized 3' end enriched library protocols such as 3' READS⁶⁷, PAS-seq⁴¹, QuantSeq⁶⁸, or TREND-seq⁶⁹ were developed. These methods selectively capture the 3' ends of polyadenylated transcripts, enabling high-resolution PAS mapping⁴. These techniques typically use oligo(dT) priming to enrich for the 3' ends of mRNA molecules, and capture PAS usage with high sensitivity. While these specialized methods significantly improved APA detection resolution, they also presented technical challenges. A common issue in identifying PAS using these methods rely on the usage of oligo(dT) primer to capture poly(A) mRNA molecules. These primers typically bind to internal adenine-rich sequence stretches rather than true poly(A) tails, and can falsely lead to the identification of these regions as potential PAS¹³.

1.1.6.2.3. Long-Read Sequencing Technologies

More recently, the development of long-read sequencing platforms, such as PacBio Iso-Seq or Oxford Nanopore, enabled unambiguous identification of APA isoforms by providing full-length transcript coverage^{70,71}. Unlike short-read sequencing, these technologies do not rely on computational reconstruction and can unambiguously identify APA isoforms^{72,73}. However, their application to large scale studies remain limited due to higher error rates (typically 2–5% for PacBio

and ~7% for Oxford Nanopore, versus ~0.005% for Illumina), low-throughput, and higher cost per read compared to short-read sequencing technologies^{74,75}.

1.1.6.2.4. Characterizing APA at Single-Cell Resolution: A New Frontier

While bulk RNA-Seq based methods have provided critical insight into APA dynamics across various biological context, their bulk-level sampling provides average signals across cell populations, masking cell-to-cell variability. The advent of scRNA-seq has addressed this limitation by enabling gene expression analysis at the resolution of individual cells^{76–78}. The inherent 3' end enrichment of most scRNA-seq platforms (e.g., 10x Genomics, DropSeq) that capture only the 3' end of transcripts generating 3' tagged based data, offers a natural advantage to estimate PAS usage and profile the isoform expression across diverse cell types or biological contexts^{38,79,80}. Nevertheless, applying APA analysis to scRNA-seq data presents unique challenges due to constraints such as the technical biases introduced related to sample preparation, low read depth per cell and dropout events, or batch effects.

1.2. Study of gene expression using single-cell transcriptomics

1.2.1. Evolution of scRNA-seq experiments

1.2.1.1. Plate-based methods

The first scRNA-seq experiment conducted by Tang et al., demonstrated the applicability of high-throughput sequencing methods at single-cell resolution by sequencing the transcriptome of a single blastomere⁸¹ (Figure 5a). This pioneering study laid the groundwork for scRNA-seq and was based on manual cell isolation into individual wells of microtiter plates, followed by full-length cDNA synthesis and amplification. While this pioneering study laid the foundation for the field, it was constrained by low-throughput and labor-intensive processing. A significant advancement in the field was the development of multiplexing strategies that enabled the analysis of multiple samples at once. The first scRNA-seq protocol that employed this approach was published Islam et al. using single-cell tagged reverse transcription sequencing (STRT-seq) technique, which used unique template-switching oligos (TSO) in each well of a 96-well plate to introduce a well-specific barcode during reverse transcription⁸². This enabled pooling of barcoded samples before sequencing, improving efficiency and reducing batch effects. Later, the combination of different multiplex approaches through the usage of multiple barcoding steps enabled an increase in the number of cells that can be sequenced in parallel with the development of scRNA-seq protocols such as MARS-seq⁸³ and STRT-Seq2i⁸⁴ (Figure 5a). For instance, within the STRT-Seq-2i protocol, up to 100 barcodes are added during the full-

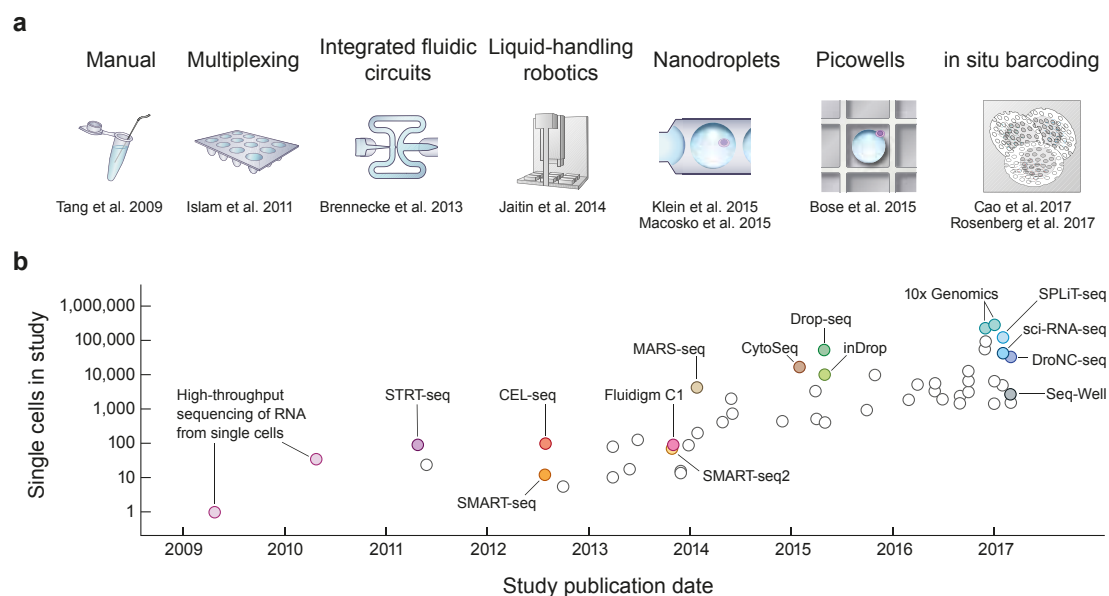


Figure 5. Evolution in the scaling of scRNA-seq experiments. **a.** Advances in experimental scale have been driven by key technological innovations. Sample multiplexing initially enabled the analysis of ~100 cells, followed by the use of integrated fluidic circuits, which increased this to ~1,000 cells. Liquid-handling robotics then facilitated studies involving several thousand cells. Subsequent orders-of-magnitude increases to tens of thousands of cells were achieved with random capture technologies using nanodroplets and picowells. More recently, in situ barcoding has allowed cost-effective analysis of hundreds of thousands of cells. **b.** Cell numbers reported in representative publications by publication date. Key technologies are indicated (adapted from Svensson et al⁸⁹).

length PCR, followed by 96 additional barcodes for different regions of their array during library preparation, enabling a total of 9,600 cells to be multiplexed and sequenced together in a single run^{76,85}. These multiplexing innovations significantly expanded the throughput and accessibility of scRNA-seq, laying the groundwork for the high-throughput platforms used today.

Subsequent developments improved both scalability and transcript coverage. Notably, the Smart-seq family of protocols brought major advances in

transcriptome resolution. Smart-seq, introduced by Ramsköld et al enabled full-length transcript profiling, addressing the limitations of 3'-biased protocols⁸⁶. This was further refined in Smart-seq2⁸⁷, which increased sensitivity and reproducibility, making it suitable for applications such as isoform detection, allelic expression, and transcript variant analysis.

1.2.1.2. Combinatorial barcoding approaches

To further scale up the number of cells that can be multiplexed, new protocols were developed using combinatorial indexing strategies. These approaches perform multiple rounds of splitting and pooling of the cells in wells to uniquely tag RNA from individual cells, increasing the total number of cells that could be sequenced in parallel. The first combinatorial approach was introduced in early 2015 with the CytoSeq method, where the authors used initial pools of 96 barcodes into which beads were split-pooled three times to generate a set of $96^3 = 884,736$ barcoded beads⁸⁸. However, Poisson statistics associated with this approach dictate that random cell-isolation methods will always be inefficient, requiring a significantly larger pool of barcodes than captured cells to prevent barcode duplication and ensure mostly single-cell isolation⁸⁹.

This limitation was addressed with the development of *in situ* barcoding protocols such as the single-cell combinatorial-indexing RNA-seq (sci-RNA-seq) by Cao et al. in 2017 or the split-pool ligation-based transcriptome sequencing (SPLiT-seq)^{90,91}. In these approaches, single cells are never individually isolated and lysed; instead, cells are fixed and the RNA is manipulated *in situ* inside each cell. Then, similar to the combinatorial indexing protocols, the cells are processed with multiple rounds of splitting and pooling into wells with the addition of a new well-

specific barcode. This procedure results in an arbitrarily low probability that any two cells will co-locate in the same sequence of wells, and so RNA from each cell is uniquely labeled, enabling to profile a large number of cells at low cost⁹².

1.2.1.3. Development of microfluidic approaches

The introduction of microfluidic technologies marked a significant advancement in the field, by enabling the handling and processing of very small volumes to encapsulate individual cells in a high-throughput manner⁹³ (Figure 5b). These platforms allowed for the isolation, lysis, and reverse transcription of RNA from single cells within compact microfluidic devices, significantly improving the efficiency and scalability of single-cell studies. Early microfluidic platforms, such as the Fluidigm C1 system, provided a means to automatically isolate and process single cells in a high-throughput manner⁹⁴. This system allowed for the sequential delivery of small volumes into tiny reaction chambers, capturing cells passively and automating several steps of the sequencing process such as RNA capture, reverse transcription, or cDNA amplification.

A further leap was achieved with the development of nanodroplet-based systems such as Drop-seq⁹⁵ and InDrop⁹⁶, and later 10x Genomics Chromium⁹⁷. These systems use the microfluidic devices to encapsulate individual cells in nanoliter-sized droplets along with barcoded beads and reagents. These methods enabled the simultaneous processing of thousands of cells, drastically increasing the cell processing throughput and reducing costs. For instance, Drop-seq, introduced in 2015 by Evan Macosko and colleagues at Harvard Medical School, revolutionized single-cell RNA sequencing by enabling high-throughput analysis of thousands of individual cells at a lower cost⁹⁵. In this technology, a suspension

of single cells is mixed with a stream of barcoded beads and a stream of lysis buffer, then flowed through a microfluidics device where they are encapsulated in droplets by an oil phase. Each droplet contains a single cell and a single barcoded bead, allowing for the capture and labeling of RNA molecules within each droplet after the lysis of the encapsulated cell. Then, barcoded RNAs are reverse transcribed into cDNA, pooled, and sequenced to reconstruct gene expression profile for thousands of single cells simultaneously.

The nanodroplet-based method became widely accessible through automation and commercialization with the introduction of 10x Genomics Chromium systems, which gradually became a primary choice for most scRNA-seq experiments (Figure 6)⁹⁷. This platform introduced the concept of gel beads-in-emulsion (GEMs), which are formed by combining 10x barcoded gel beads, partitioning oil, and a master mix containing the cells, oligonucleotides, and reverse transcription

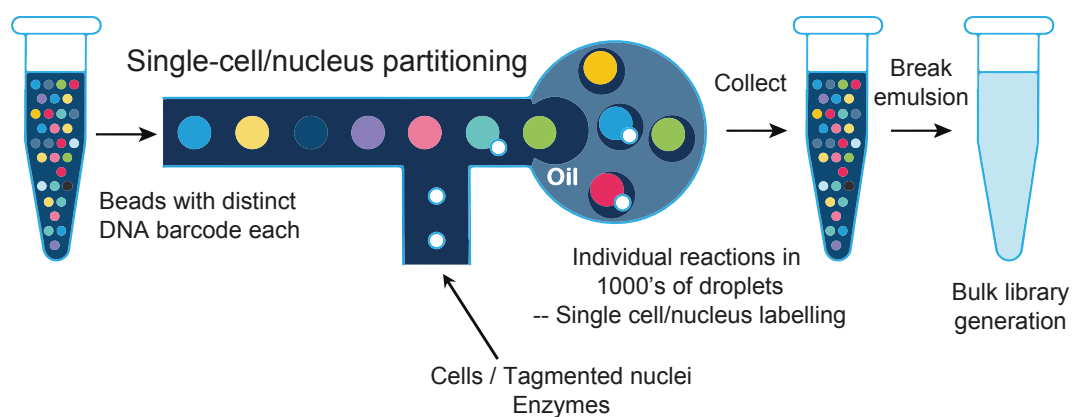


Figure 6. 10x single-cell encapsulation. Single-cell sequencing workflow for droplet-based approaches exemplified by the 10x Genomics Chromium platform: Cells are dissociated into a single-cell suspension and loaded onto a microfluidic chip, where they are encapsulated into nanoliter-scale Gel Beads-in-emulsion (GEMs) droplets with barcoded beads and reverse-transcription (RT) reagents (adapted from <https://www.10xgenomics.com/technology/>).

reagents. Within GEMs, cells are lysed and release their RNA, which are then captured by the barcoded oligonucleotides. The captured RNA undergoes reverse transcription to produce cDNA, which is then amplified. The resulting cDNA libraries, which include unique molecular identifiers (UMIs), are prepared for a high-throughput sequencing platform such as Illumina. This system enabled high-throughput scRNA-seq with streamlined workflows, supporting tens of thousands of cells in a single experiment.

1.2.2. Basic scRNA-seq workflow

Current scRNA-seq workflows, particularly those based on droplet microfluidics (e.g., 10x Genomics Chromium, DropSeq), have standardized a robust and scalable pipeline for profiling thousands of cells in parallel. Despite methodological variations, most scRNA-seq protocols follow five core steps (Figure 7):

- **Sample preparation:**

Preparation of high-quality single-cell suspensions is key to successful single-cell studies. Irrespective of the starting material, the condition of the cells is critical for efficient cell capture and optimal performance of the scRNA-seq protocols⁹⁸. Although most methods use fresh viable single cells, alternatives include preserved samples^{99–101} or cell nuclei^{102,103}. Cell isolation strategies range from manual pipetting and laser capture microdissection to high-throughput

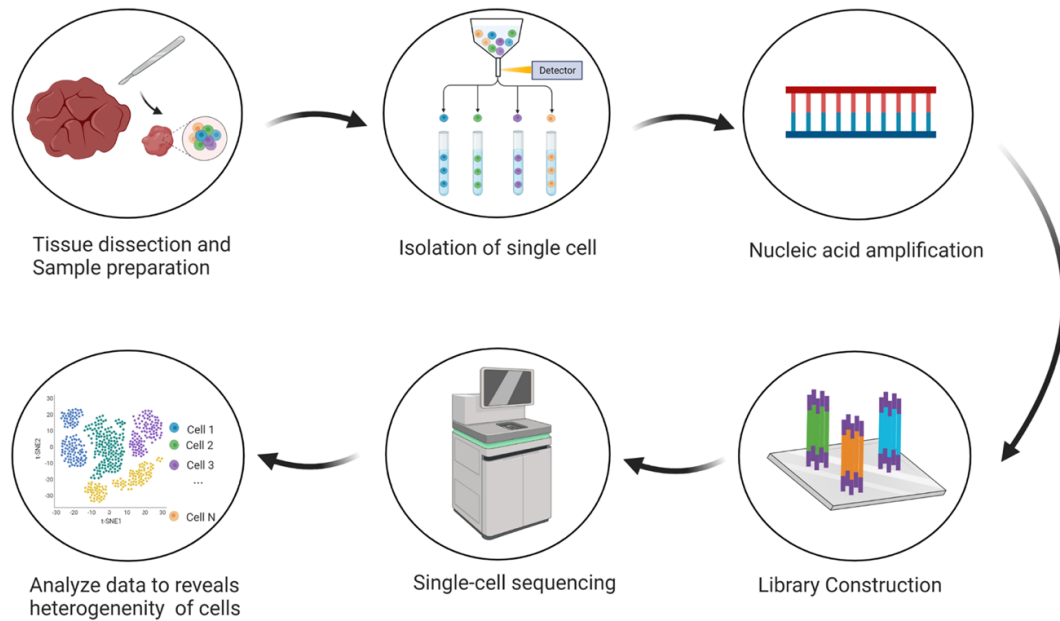


Figure 7. Main steps in single-cell transcriptomics. the main process of single-cell sequencing technology includes sample preparation, single-cell isolation, nucleic acid amplification, library construction, single-cell sequencing, and data analysis (adapted from Qu et al²⁵⁵)

fluorescence-activated cell sorting (FACS)¹⁰⁴ and automated microfluidic encapsulation into droplets or nanowells¹⁰⁵.

- **RNA capture and reverse transcription:**

Once isolated, mRNA from each cell is captured and reverse transcribed into complementary DNA (cDNA). This step typically involves the use of oligo(dT) primers to selectively capture polyadenylated mRNAs^{81,82}, and the usage of unique barcode tags to label the RNAs from each cell. Several protocols such as Drop-seq or 10x Genomics additionally incorporate Unique Molecular Identifiers (UMIs) into each cDNA molecule to mitigate amplification bias and improve transcript quantification accuracy^{95,97}.

- **cDNA amplification:**

To generate sufficient material for sequencing, the cDNA is amplified typically using a polymerase chain reaction (PCR) or through *in vitro* transcription (IVT)^{85,87}.

- **Library construction and sequencing:**

Amplified cDNA is enzymatically fragmented and ligated with platform-specific sequencing adapters. Libraries are sequenced using high-throughput sequencing platforms such as Illumina or Ion Torrent, producing millions of read sequences per run^{106,107}. The depth of sequencing required depends on the research question and the number of cells analyzed, with deeper sequencing providing more comprehensive transcriptomes¹⁰⁸.

- **Data processing and analysis:**

Finally, sequencing reads are demultiplexed, aligned to a reference transcriptome, and collapsed into a digital gene-cell count matrix (DGE). Computational workflows perform normalization and batch correction (e.g., SCTransform, Harmony), dimensionality reduction (e.g., PCA, UMAP, t-SNE) to visualize high-dimensional single-cell data^{109,110}, clustering, and differential expression analysis to define cell states and marker genes^{111,112}. Additional analyses, such as trajectory inference and pseudotime analysis, can be applied to explore cell lineage relationships and dynamic gene expression changes^{113,114}.

1.2.3. Library Construction Strategies

Single-cell libraries are typically generated using two main strategies: full-length and tag-based methods^{76,115}. While the full-length methods such as Smart-seq1/2/3^{86,87,116}, SUPeR-seq¹¹⁷, or MATQ-seq¹¹⁸ produce data that covers the whole transcript, tagged based methods only capture and sequence the 3' end of RNAs (Drop-seq, Seq-Well, SPLiT-seq, Chromium Single Cell 3' Gene Expression) or 5' end (STRT-seq, Chromium Single Cell 5' Gene Expression), introducing a bias in the distribution of reads along transcripts towards the 3' or

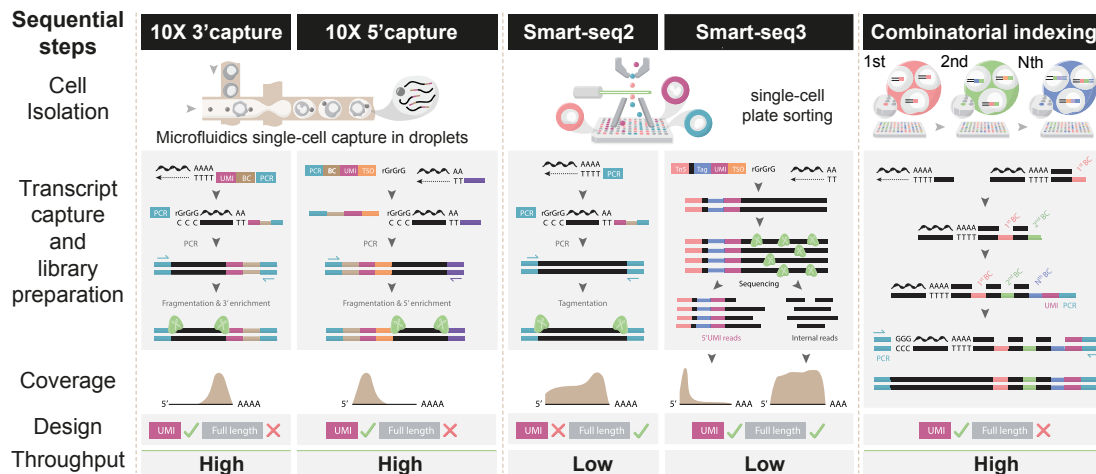


Figure 8. Overview of main scRNA-seq library preparation approaches. Most single-cell protocols capture mRNA poly(A) tails using oligo-dT primers followed by reverse transcription. A key difference between the tagged-end methods, such as 3' and 5' capture protocols from 10x Genomics and Smart-seq methods, lies in the targeted amplification of the tagged ends after fragmentation/tagmentation rather than full-length amplification. In addition, Smart-seq3 incorporates UMIs at the 5' end of transcripts. Both 5' UMI reads and internal non-UMI reads may be strategically combined to improve transcript quantification. Combinatorial barcoding methods use both oligo-dT and random hexamers for transcript capture. The generation of unique cell barcodes relies on an initial in situ retro-transcription step followed by several rounds of in situ ligation and posterior RT + template switching (adapted from Conte et al ⁷⁸).

the 5' end regions (Figure 8). 3' end tagged-based methods specifically target and enrich sequences at the polyadenylated 3' end of mature mRNA transcripts using oligo-dT primers, followed by high-throughput sequencing. Each primer includes a UMI and a cell-specific barcode tag, enabling precise quantification and identification of mRNA molecules from each cell, whereas 5' end tagged-based methods capture the 5' ends of mRNAs using primers complementary to the 5' cap.

In contrast to tag-based methods, full-length protocols such as Smart-seq library approaches provide a full-length transcript coverage. Although prior methods such Smart-seq 1/2 did not include UMI tag, making molecule level quantification less accurate, Smart-seq3 addressed this limitation by introducing UMI at 5' end of transcripts, thereby enabling full-length coverage and improved quantification accuracy¹¹⁶. However, these plate-based approaches have a lower-throughput compared to droplet-based methods, and are associated with higher experimental cost per cell^{119,120}.

1.2.4. Sequencing methods

The sequencing methods can be broadly categorized into two categories: 1) short-read sequencing and 2) long-read sequencing. Short-read sequencing platforms such as Illumina¹²¹ are widely used in single-cell RNA-seq due to their high accuracy and scalability, generating reads of ~100–150 bp typically from either the 3', 5', or full-length transcripts depending on the library preparation method. In contrast, long-read sequencing technologies such as SMRT, or MinION capture an entire transcript molecule in a single sequencing read, enabling a full-length sequencing of the isoforms and their quantification in an

unambiguous manner^{122,123}. However, these methods typically show higher error rates than short-read technologies that can be problematic for single-cell RNA-seq isoform studies. For instance, typical error rates for PacBio sequences are in the range of ~2–5%¹²⁴, whereas for Oxford Nanopore, the values go up to ~7%^{125,126}. This contrasts with the high accuracy of Illumina sequencing (~0.005% error rate). Additionally, these techniques generally have lower-throughput, limiting the number of cells analyzed in large-scale studies¹²⁷.

1.3. Challenges in the study of gene expression in single-cell data

1.3.1. Dropout and batch effects

Due to scarcity of RNA material in a single cell and the nature of stochastic transcription, scRNA-seq data often faces challenges such as high cell-to-cell variability and dropout effects^{77,78,121,128}. Such variability is caused by both technical noise (e.g. RNA capture efficiency, random dropouts during library preparation) and biological noise (e.g. stochastic gene expression, various cellular states, cell sizes, and cell cycle state). These dropout events occur due to the low amounts of mRNA in individual cells and inefficient mRNA capture, as well as the stochasticity of mRNA expression. As a result, scRNA-seq data are often highly sparse, and excessive zero counts cause the data to be zero-inflated, only capturing a small fraction of the transcriptome of each cell and affect the accuracy of quantification of APA.

Additionally, experimental batch effects are frequently found in high-throughput data^{129,130}. Batch effects can arise from various causes, such as differences in reagent lots, library preparation protocols, sequencing platform, or sample handling procedure across experiments. While careful experimental design and the use of multiple biological replicates can minimize batch effects, the variability in gene detection rates due to variable capture efficiency and genetic background of biological specimens remains difficult to control by experimental procedures. Multiple batch correction methods have been developed to deal with these issues in single-cell datasets such as Harmony¹³¹, Seurat¹³² or MMD-ResNet¹³³. These methods enable processing scRNA-seq dataset from different batches by

removing technical variation while preserving the biological variability. Thus, the correction of batch-specific differences allows for accurate clustering, differential gene expression analysis and cross-sample comparisons¹³⁴.

1.3.2. Impact of cell dissociation on single-cell data

Sample preparation is a critical upstream component of scRNA-seq workflows and can significantly influence the quality and interpretability of the resulting data. Key steps such as tissue dissociation, cell isolation, cell lysis, and RNA preservation represent arguably a great source of technical variation, contamination, and batch effects in any single-cell study^{135–138}, and hence represent a key determinant for a successful experimental design. Dissociation of the cells is highly tissue-specific and typically involves the use of different strategies including enzymatic digestion (e.g., collagenase, trypsin, liberase), mechanical disaggregation (e.g., pipetting, filtering), fluorescence-activated cell sorting (FACS), and other technologies to generate viable single-cell suspension. This step has been shown to introduce substantial technical artifacts, including transcriptional stress responses, selective loss of fragile cell types, or altered gene expression profiles that do not reflect the *in vivo* state. For example, warm digestion at 37 °C can activate immediate-early genes such as *FOS*, *JUN*, and *EGR1*, as well as stress and inflammatory pathways, which may obscure true *in vivo* gene expression profiles^{135,139}.

1.3.3. Impact of sample preservation methods on the gene expression

Initial scRNA-seq protocols were designed to work optimally with freshly isolated cells to ensure high-quality transcriptomic data. However, in both research and clinical context, immediate sample processing is not always feasible due to logistical constraints^{98,140}. Moreover, processing samples at different time points can introduce batch effects and interfere with downstream analysis. To address these issues, various sample preservation techniques can be employed to decouple the sample collection from processing, and enable the storage and processing of the samples on different time points while minimizing technical variability¹⁴¹.

1.3.3.1. Overview of preservation methods and protocols

Cryopreservation is one of the most widely used methods, relying on cryoprotective agents such as the dimethyl sulfoxide (DMSO) to freeze cells at ultra-low temperatures^{101,142}. DMSO helps prevent cellular damage by preventing the ice crystal formation during freezing process, and enable long-term storage of cells and tissues¹⁴². However, this method may introduce stress-related responses during thawing process, especially in sensitive cell populations¹³⁵.

In contrast to DMSO, chemical fixation methods offer the advantages of sample preservation at room storage temperature. Main protocols include methods such as methanol fixation^{99,143}, acetic-methanol (ACME)¹³⁶ paraformaldehyde (PFA)¹⁴⁴, or dithio-bis-succinimidyl propionate (DSP) also known as Lomant's reagent¹⁴⁵. With the methanol fixation, cells are permeabilized and stabilized

through dehydration, enabling them to be stored at 80° C until processing, making it a practical option for high-throughput scRNA-seq, although fixed cells tend to lose their morphological information¹⁴⁶. PFA, on the other hand, is effective in preserving both RNA content and cell structure integrity by cross-linking the proteins and nucleic acids¹⁴⁴. Similarly, DSP maintains both cellular structure and RNA integrity, while offering reversible crosslinking for RNA recovery post-fixation¹⁴⁵.

Moreover, commercial preservation solutions, such as CellCover¹⁴², and vivoPHIX¹⁴⁷, provide convenient alternatives. These formulations are designed for room temperature use, and claim to maintain both transcriptomic integrity and cellular composition without requiring conventional freezing or fixation setups¹⁴⁸.

1.3.3.2. Effect on transcriptomic quality and cell composition

Although sample preservation techniques enhance the practicality of scRNA-seq experiments by enabling long-term storage and delayed processing, multiple studies highlighted that preservation procedures introduce variability on the samples that impact both gene expression profiles and cellular composition^{99,101,143,145}. These effects vary from uneven preservation of RNA integrity, and biases in transcript detection, especially for genes expressed at low levels^{99,142}. For instance, Guillaumet-Adkins et al reported that cryopreservation of human colon tumor tissues using DMSO led to a decreased detection rate of low-expression genes, despite maintaining overall transcriptome complexity¹⁰¹.

Additionally, transcript dropout is another major issue introduced by preservation, particularly for genes with high turnover. Denisenko et al conducted a

comparative study of several preservation protocols, and showed that chemical fixation, such as methanol or PFA, increased the dropout rate, thereby skewing the expression landscape¹³⁵. This bias can lead to the underrepresentation of regulatory or signaling genes, which, although lowly expressed can be crucial for interpreting cell state dynamics. Furthermore, sample preservation can lead to cell compositional biases by disproportionately affecting specific cell type populations¹³⁵. Fragile or stress-sensitive cell types such as certain neuronal subpopulations, epithelial cells, or immune cells show higher vulnerability due to the physical and chemical stresses introduced by freezing or fixation^{99,135}. Alles et al., for example, noted a reduced recovery of specific peripheral blood mononuclear cell (PBMC) subtypes following cryopreservation⁹⁹. Similarly, Jaitin et al. showed that delays between tissue collection and processing could shift the balance between immune cell types, such as T cells and monocytes⁸³. Such changes can mislead biological interpretation, particularly when comparing fresh versus preserved samples or across clinical cohorts.

Taken together, while preservation methods are invaluable for expanding the logistical feasibility of scRNA-seq studies, they introduce biases that pose significant hurdles for accurate gene expression analysis. These challenges are particularly acute when attempting to detect isoform usage and unravel APA dynamic at single-cell resolution. This problematic poses the need for careful benchmarking to mitigate preservation-induced biases and the development of dedicated computational methods to fully leverage the scRNA-seq data for APA study.

1.4. Characterization of APA at single-cell resolution

1.4.1. Emergence of methods for APA profiling from single-cell data

The advent of scRNA-seq opened new opportunities for characterizing APA at single-cell resolution. Although early scRNA-seq protocols were not specifically designed to study APA, their inherent bias toward the 3' end of the transcripts provided a foundation to profile PAS usage across different cell types. In particular, widely adopted droplet and 3' tagged-based scRNA-seq protocols, such as Drop-seq⁹⁵, and 10x Chromium⁹⁷ offered strong potential for dissecting APA in single cells. However, the extremely high dropout rate and cell-to-cell variability inherent in scRNA-seq make it challenging to directly apply bulk RNA-seq methods developed for APA characterization on scRNA-seq data¹⁴⁹. Consequently, recent years have seen the development of computational approaches specifically designed for APA analysis at single-cell resolution and employing diverse strategies to address the unique challenges related to scRNA-seq data.

1.4.2. Computational methods for APA characterization

Several computational methods have been developed to identify and quantify APA in scRNA-seq data^{79,80,150–152}. Most methods rely primarily on genome reference annotations, while others incorporate scRNA-seq data along public PAS databases. According to their implementation strategies, these methods could be globally clustered into two categories: peak calling-based methods and density distribution-based methods¹⁵³.

1.4.2.1. Peak Calling based methods for APA characterization

The peak calling strategy is widely used by most computational methods for the analysis of APA using scRNA-seq¹⁵³. The core principle of this method consists in the identification of areas in the genomic landscape for which the aligned reads from 3' tagged-based scRNA-seq accumulate to form peaks susceptible to be identified as potential PAS (Figure 9). Next, the identified peaks are filtered based on their read enrichment in genomic regions, and annotated to distinguish true APA events from background noise or technical artifacts^{150,151}.

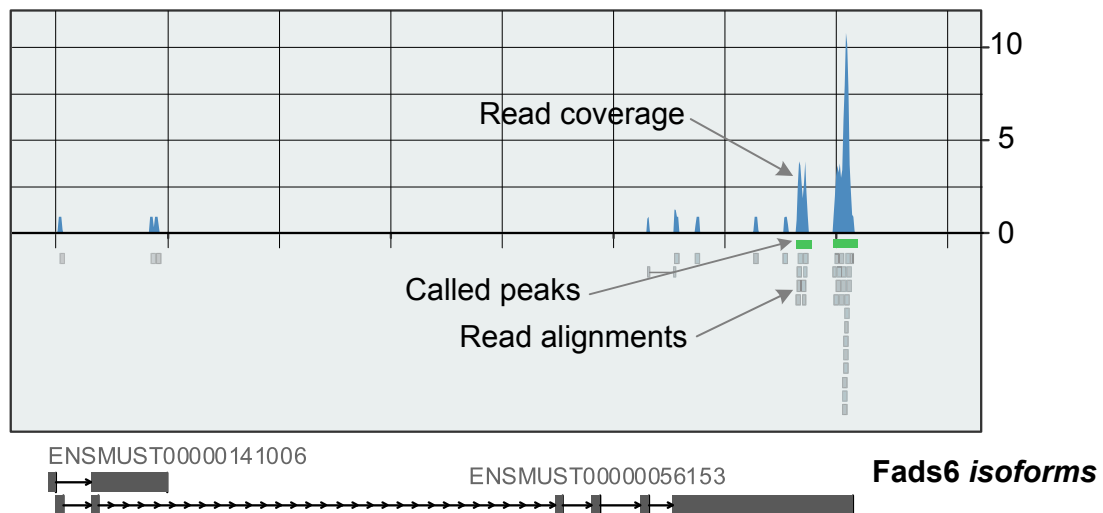


Figure 9. Illustration of the peak calling based methods for PAS identification using 3' enriched scRNA-seq data. Low coverage peaks are missed by peak calling-based methods and overlapping peaks resulted from usage from adjacent PAS cannot be separated (adapted from Ye et al. ¹⁵³).

This approach is implemented in varying ways across several tools (references). For instance, scAPAtap performs a genome-wide peak calling and poly(A) read anchoring to accurately identify PAS without relying to prior genome annotation¹⁵¹, while other tools restrict the search to specific genomic features such as 3' UTR or exonic domains, enabling to detect distinct types of APA

events^{38,153}. For instance, in scAPA only non-overlapping 3' UTRs are used for peak calling using HOMER^{38,154}. To handle adjacent PAS, scAPA applies a Gaussian finite mixture model to split larger peaks detected into smaller ones. In contrast, Sierra uses splice-aware peak calling approach on the whole genomic space, and applies a Gaussian curve fitting to identify potential PAS peaks, which are then annotated and quantified in individual cells¹⁵⁰. In these methods, the peak calling process is performed on the genomic space, thereby does not allow to link the identified PAS to specific isoforms. SCAPTURE, developed by Li et al. performs a peak calling at the transcript level using genomic annotation and publicly annotated PAS database, enabling to capture PAS information from scRNA-seq data and assign them to overlapped transcripts. Although peak calling based methods can rely on public PAS database to identify and quantify high-confidence PAS, a central advantage lies in their ability to discover previously unannotated PAS, thereby extending APA profiling beyond known cleavage sites.

However, peak-calling based approaches show significant limitations, particularly due to the high dropout rate and sparse and uneven read coverage inherent to scRNA-seq datasets (Figure 9). These factors introduce challenges in distinguishing true PAS from technical noise, increasing the risk of biases related to false-positive identification¹⁴⁹. Additionally, although peak calling based methods identify genomic locations that can be considered as PAS, they cannot reliably link these sites to a specific isoform. This is due to the fact that a single genomic location can overlap multiple isoforms, making it difficult to assign the cleavage site to a particular transcript. As a result, this approach is not suitable

for investigating differential isoform usage across cell-types or identifying how specific isoforms might be preferentially expressed under certain conditions⁷.

1.4.2.2. Distribution-based methods

Distribution-based methods characterize APA events by leveraging the read distribution in the genomic space. Instead of identifying discrete peaks, distribution-based approaches model the entire read coverage profile, using statistical or probabilistic frameworks to detect shifts in coverage patterns across transcript or genomic regions that could be linked to APA events (Figure 10)¹⁵³. Rather than calling precise locations and boundaries of peaks, these methods quantify the differential relative usage between groups of peaks to overcome the fuzzy distribution of the reads in the genomic space. During the course of my PhD study over the last 4 years, several computational methods implemented for APA characterization in single-cell data utilizing distribution-based approaches have been released^{79,151,152,155}.

For instance, MAAPER implemented in 2021 by Li et al., uses a probabilistic model to assign scRNA-seq reads to known PAS, estimating their usage while accounting for uncertainty in read assignment¹⁵⁵. Another example is scDapars, which was developed to handle the specific challenges posed by scRNA-seq

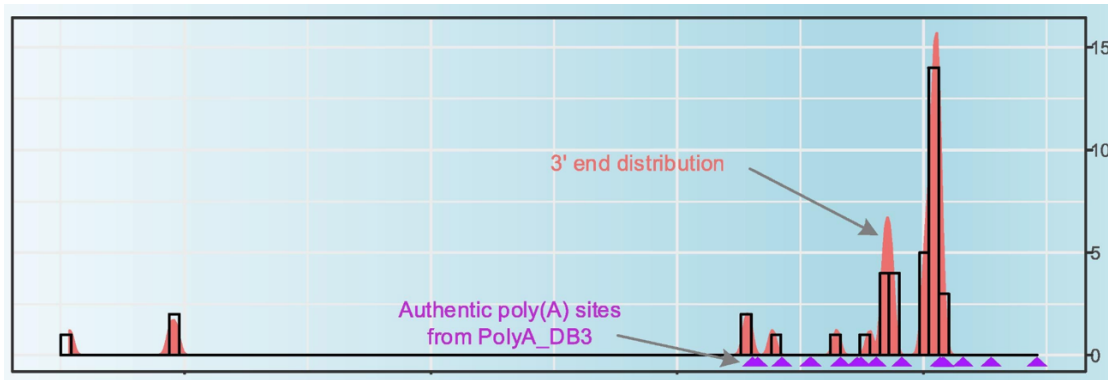


Figure 10. Illustration of the distribution-based methods for PAS identification using 3' enriched scRNA-seq data. These methods utilize statistical methods to model the complete read coverage profiles across genomic regions, and detect shifts in read coverage that are indicative of APA events (adapted from Ye et al.¹⁵³).

data, such as sparse coverage and high dropout rates. scDapars employs a regression-based model to infer the dynamic usage of proximal and distal PAS across cell types or conditions by examining the read coverage distribution along 3' UTRs⁵³ at single-cell level⁸⁰. Recently in 2024, Fansler et al., published scUTRquant, a new tool that quantifies 3' UTR isoforms from scRNA-seq data⁷⁹. This method relies on Kallisto bustools¹⁵⁶ to perform pseudo-alignments of the reads on a truncated set of reference transcripts and proportionally assign ambiguous reads to transcripts based on their distribution. This isoform quantification enables the detection of alternative 3' UTR isoform usage that could result from APA events. However, scUTRquant primarily relies on a pre-filtered cleavage site atlas established from GENCODE annotation (internal priming filtering / PAS filtering) and a Microwell-seq dataset of 400,000 single cells for human and mouse organisms, limiting the detection of novel isoforms and/or its application to late-released genome references.

2. Objectives

The main aim of this thesis is to **develop a novel computational approach to characterize the isoform diversity at single-cell resolution using 3' tagged scRNA-seq data** to enhance the understanding of APA dynamics across cell types. To achieve this aim, we will pursue the following goals:

2.1. Objective 1: Optimize single-cell transcriptomics protocols to work with iPSC-derived neuron and glial cells

1. Assessment of the impact of sample preservation methods on the library complexity and cell composition of scRNA-seq datasets
2. Investigation of the effect of sample preservation methods on the gene expression profile of individual cells

2.2. Objective 2: Implementation of a new computational method for the quantification of 3' end isoforms using scRNA-seq datasets

1. Development and implementation of SCALPEL, a new pipeline for isoform quantification at single-cell resolution
2. Benchmarking of SCALPEL performance against published tools
3. Usage of single-cell 3' end isoform quantification data enable to identify of cell populations
4. Investigation of APA and 3' end isoform variability across the cell types

3. Material and Methods

3.1. Objective 1: Optimize single-cell transcriptomics protocols to work with iPSC-derived neuron and glial cells

3.1.1. Sample preservation methods

We prepared single-cell suspensions from iPSC-derived neural progenitor cells by enzymatic dissociation¹⁵⁷ using a 1:1 mixture of papain and accutase, followed by filtering through a 40 μm strainer and viability assessment by trypan blue exclusion¹⁵⁸. We applied five different preservation methods to hiPSC-derived neural progenitor cells. For DMSO cryopreservation, we suspended 2.5×10^6 cells in neural maintenance medium (1:1 ratio of DMEM/F-12 GlutaMAX (Gibco, #10565018)) with 10% v/v DMSO and 20 ng ml^{-1} bFGF, froze them overnight at -80°C in a Mr. Frosty container, and stored them in liquid nitrogen. For thawing, we used a 37°C water bath, diluted the cells in medium, centrifuged them, and washed the pellet in DPBS with 0.01% BSA. For methanol fixation, we resuspended 2.5×10^6 cells in ice-cold DPBS, added chilled methanol dropwise to reach 80%, incubated the cells on ice for 30 min, stored them at -80°C . Rehydration of methanol-fixed samples was done by the addition of DPBS with 0.01% BSA, $0.2 \text{ U } \mu\text{l}^{-1}$ of RNase inhibitor, and 1 mM DTT. For ACME fixation, we added ACME solution (13:3:2:2 mix of wash buffer (DPBS with 0.01% BSA and $0.2 \text{ U } \mu\text{l}^{-1}$ of RNase inhibitor and 1 mM DTT), methanol, acetic acid, and glycerol) to a pellet of 1 to 5×10^6 cells, incubated for 30 min at room temperature, washed twice in wash buffer, and stored at -80°C in wash buffer with 10% v/v DMSO. We rehydrated ACME-fixed samples using the same procedure used for methanol-fixed samples. For vivoPHIX fixation, a cell pellet containing 1 to 5×10^6 of cells was resuspended with 25 μl of DPBS with 0.01% BSA and fixed by

the addition of 75 μ l, of vivoPHIX reagent (Rapid Labs) (3:1 ratio), mixed for 5 min, rotated at room temperature for 30 min, stored at 4 °C overnight, and transferred to -80 °C. For rehydration, we added one volume of 100% ethanol, mix and centrifuged, treated with vivoPHIX-SCAA (1:3 vivoPHIX:acetic acid) for 3 min, and washed the pellet with DPBS/BSA/RNase inhibitor/DTT before filtering as it was done for the other fixed samples. For CellCover fixation (Anacyte Laboratories), we added 10 volumes of CellCover to $1-5 \times 10^6$ cells and stored them at 4 °C or 80 °C. Due to strong degradation level with RNA integrity value ~ 2 at 4° C and ~ 6 at 80 °C, we discarded it as not suitable for long term cell storage.

3.1.2. Single-cell RNA-seq library preparation

For single-cell encapsulation, we included only samples with viability above 75%. We loaded 75,000 cells in 250 μ l of PBS with 0.01% BSA and 150,000 washed Macosko barcoded beads in 250 μ l of lysis buffer (6% Ficoll PM-400, 0.2% Sarkosyl, 0.02 M EDTA, 0.2 M Tris-HCl pH 7.5, and 0.05 M DTT) into the NADIA microfluidic instrument (Dolomite Bio, now Unchained Lab) and co-encapsulated them in droplets using standard settings, achieving a capture efficiency of 5–7%. We immediately broke the emulsion and reverse transcribed captured RNAs on beads using Maxima H Reverse Transcriptase Master Mix. We removed excess primers by incubating beads with Exonuclease I at 37 °C for 45 minutes. We washed and resuspended beads in nuclease-free water at 400 beads μ l⁻¹, split them into pools of 4000 beads per PCR tube, and amplified the cDNA using 9 or 11 PCR cycles depending on the bead batch. We purified cDNA with 0.6 $\times$ AMPure XP beads, quantified it using Qubit dsDNA HS Assay, and assessed size profiles with an Agilent 4200 TapeStation. For library construction, we used 600 pg of

cDNA per sample and the Nextera XT DNA Library Prep Kit (Illumina), followed by adapter tagging, PCR amplification, and purification with AMPure XP beads. We quantified and verified library size with Qubit and TapeStation, pooled the libraries, and sequenced them on an Illumina NextSeq 550 platform using the High Output v2 kit (75 cycles) in paired-end mode: 20 bp for Read 1 using the custom Read1CustSeqB primer⁹⁵ (capturing cell barcode and UMI), 64 bp for Read 2, and 8 bp for the i7 index.

3.1.3. scRNA-seq data pre-processing

The processing of scRNA-seq libraries was performed using the Drop-seq tools 2.3 pipeline^{95,159} to generate Digital Gene Expression (DGE) matrices. Index and annotation reference files for the human genome (hg38 assembly) were generated using Ensembl v100 annotation¹⁶⁰. Next, paired-end FASTQ files were merged into a single unaligned BAM file using Picard tools v2.18.14¹⁶¹. Using the Drop-seq toolkit with default settings, reads were tagged with the cell and the molecular barcodes, trimmed at their 5' and 3' end to remove the adapter sequences and poly(A) tails. Using STAR version 2.7.0.a (Dobin et al., 2013), reads were mapped to the human genome (version hg38). Resulting aligned bam files were tagged with the annotation metadata files to identify reads overlapping genes. Finally, cell barcodes were corrected using *DetectBeadSubstitutionError* and *DetectBeadSynthesisErrors* from Drop-seq tools with default parameters. A Knee plot was generated to estimate the number of cells obtained during the single-cell encapsulation. The number of uniquely mapped reads assigned to the top N barcodes was used as input, with N being at least five times the number of expected cells. Using the estimated number of cells obtained with this procedure,

two DGEs were generated. First DGE was generated with all the UMIs overlapping gene regions using the parameters `LOCUS_FUNCTION_LIST = INTRONIC`, `LOCUS_FUNCTION_LIST = INTERGENIC` and another one containing all UMIs overlapping introns using the parameters `LOCUS_FUNCTION_LIST = null` and `LOCUS_FUNCTION_LIST = INTRONIC`.

3.1.4. Filtering of low-quality cells and doublets

DGE expression matrices were analyzed using Seurat v 4.2.1¹⁶². Distinct Seurat objects were first created for each dataset and merged before filtering of low quality cells. Next, all the cells with a UMI count below 200 or above 17000, a gene number below 200 or above 5500, a percentage of mitochondrial transcripts higher than 7.5%, and a ribosomal content higher than 40% were discarded. The number of cells discarded at each step is reported in **Supplementary table 1**. Doublet cells were removed using DoubletFinder¹⁶³ on each sample. The parameters and doublets identified in each dataset are detailed in **Supplementary table 2**. Individual Seurat objects were merged to perform a joint analysis. All genes expressed in less than three cells were removed. A linear model was fitted to describe the relationship between the log number of UMIs and the log number of genes detected per cell. All cells with a residual below -0.5 (3 cells) were discarded. The final Seurat object obtained contained 16,870 cells and 24,468 genes.

3.1.5. Identification of cell populations

We regress the percentage of mitochondrial transcripts, the number of genes, the number of UMIs, and the preservation method were regressed within the data

using the Seurat function *ScaleData*. Next, we normalized the data using *LogNormalize* function using a scale factor of 10,000. The top 2,000 variable genes were used to compute 100 principal components (PCs). Based on the *ElbowPlot*, we selected the first 20 PCs to build the kNN graph and generate the UMAP (500 training epochs). Batch effects were removed using Harmony¹³¹. *RunHarmony* was applied with sample ID as the integration variable. By inspecting the updated Elbow plot, we selected the first 19 corrected PCs to perform the clustering. We used the package *clustree*¹⁶⁴ to inspect the clustering results at different resolutions from 0.1 to 1 and chose a final resolution of resolution 0.7 where we obtained 12 cell populations. To calculate the top markers for each cluster, we used the function *FindAllMarkers* from Seurat. We restricted the analysis to positive markers and keep all other parameters at their default settings. Statistically significant markers with an FDR-adjusted pvalue smaller than 0.05 using Wilcoxon rank-sum test were selected.

3.1.6. Stress and apoptosis gene signatures

We used the function *AddModuleScore* with default parameters from Seurat package¹⁶² to assess if fixation methods induced stress or apoptosis. This function compares the expression of a set of given genes with random sets of genes with similar expression in the dataset to calculate an enrichment. For the apoptosis signature, we built a gene signature including the genes BCL2, TNF, TP53, CASP3, BAX, CASP8, FAS45. For the stress signature, we used the following immediate early genes FOS, JUN, EGR1, UBC, HSPA1B, BTG2, IER2, ID330. Statistically significant differences in the signature score between the

different fixation methods and fresh samples was calculated using a one-tailed Wilcoxon rank-sum test (pvalue < 0.05, FDR adjusted).

3.1.7. Cell composition analysis

To assess changes in cell composition we used scCODA ¹⁶⁵. To run scCODA we defined fresh samples as a reference condition. To set the comparison, a cluster with low variability across samples was chosen as a reference. In this case, we used as reference the NPC cluster, which had a good number of cells and a very low amount of dispersion (expressed as differences between groups). To ensure the results were consistent and reproducible we ran scCODA ten times using the Hamiltonian Monte Carlo sampling method with default parameters and averaged the results. Cell types with average scores below (or above) zero have a significant decrease (or increase) in abundance according to scCODA model (false discovery rate <0.05).

3.1.8. Differential expression analysis

We used the function *aggregateData* from the R package Muscat ¹⁶⁶ to obtain pseudobulk expression values for each cluster in each of the samples. Then, we used the function *pbDS* to perform differential gene expression analysis using DESeq2 ¹⁶⁷ for each cluster and identify DEGs for each method in comparison to fresh samples. Statistical test was an FDR adjusted Wald test p-value <0.05 with an absolute log2 fold change > 0.58 Figure 15c. To assess if the different fixation/preservation methods introduced biases in the expression of particular gene sets, we investigated the biological processes associated to up and downregulated genes using the *enrichr* function of the GSEAPy package ¹⁶⁸. For

this analysis, we used all genes that were consistently up- or downregulated across at least four cell types for each fixation/preservation method. Only gene sets with at least ten genes were analyzed. As background set for the enrichment analysis, we provided all genes expressed in the dataset which had at least 445 UMIs, which is the minimum expression among the DEGs identified. Across all comparisons, we did not identify any significantly overrepresented gene ontology term (hypergeometric test adjusted P-value <0.001).

3.1.9. Correlation analysis

To assess the correlation across datasets, we compared the expression profiles of all cells from each dataset globally and at the cluster level. For that purpose, we computed pseudobulk expression values for all genes within a specific cluster/dataset. Afterwards, we log transformed the counts c using a pseudocount so that the normalized expression n was $n = \log(c + 1)$. We used these normalized expression values to calculate the Pearson correlation coefficient per cell type/sample using *cor* function in R. We used the function *pheatmap* from the R package *pheatmap*¹⁶⁹ to perform a hierarchical clustering using a complete clustering method and using the correlation coefficients to calculate Euclidean distances between clusters. Given that many cell clusters are very similar, we repeated the same procedure for each cell type separately and determined the statistical significance of sample clusters using the *shc* function from the *sigclust2* R package¹⁷⁰ on the corresponding correlation matrix. The *shc* method uses a Monte Carlo simulation-based significance testing procedure to assess the significance of the hierarchical clustering results of the dataset. Statistical significance is evaluated at each node along the hierarchical tree (dendrogram)

starting from the root using a gaussian null hypothesis test, and a corresponding p-value is calculated using the 2-means cluster index, a statistic sensitive to the null and alternative hypotheses. A family-wise error rate controlling the procedure is applied to correct for multiple testing. We generated dendrogram plots for each cell type to illustrate the similarity among clusters for different samples and highlight the statistically significant differences.

3.2. Objective 2: Implementation of a new computational method for the quantification of 3' end isoforms using scRNA-seq datasets

3.2.1. Annotation preprocessing

3' based scRNA-seq protocols use oligo(dT)s to capture polyadenylated RNAs, which introduces a bias in the location of reads towards the 3' end of the RNAs (Supplementary figure 1). Thus, we truncated the annotated isoforms in the existing annotation and restricted the quantification to isoforms that are different at the 3' end. Using GENCODE¹⁷¹ annotation as reference, we truncated all isoforms to include the last 600 nucleotides of spliced sequence from their 3' end, which is the region that displays coverage by the 3' tag-based scRNA-seq data (Supplementary figure 1). Then, we collapsed truncated isoforms with exact intron/exon boundaries and fewer than 30 nucleotides differences in their 3' end coordinates. When multiple isoforms were collapsed, we kept the name of the isoform having the higher expression according to pseudobulk quantification. For this purpose, we used *salmon quant* v1.10.3¹⁷² with default parameters to quantify all isoforms in bulk using as input the scRNA-seq bam files provided. For the analysis of multiple samples, isoform expression was averaged across all samples.

3.2.2. Read preprocessing

We processed all the input BAM files containing aligned and tagged scRNA-seq reads to discard artifacts and reads not supporting annotated transcripts. First, we used *samtools* v1.19.2¹⁷³ to split the input BAM files by chromosome using

the command *samtools view* and converted them into BED files using *bam2bed* command from BEDOPS v2.4.41¹⁷⁴. We included all reads in the bed file (option *–all-reads*) and split them into separate entries (option *–split*) if contained Ns in the CIGAR line (*i.e.* spliced reads). We used the function *findOverlaps* from GenomicRanges R package v1.50.0¹⁷⁵ to overlap the reads with the set of selected isoforms using default parameters. Given that all reads with the same BC and unique molecular identifier UMI are likely generated from the same original RNA molecule, we grouped them into a unique fragment and jointly evaluated them during isoform quantification (Supplementary figure 2). We defined the genomic coordinates of each fragment as the most 5' and 3' coordinates of its associated reads. We discarded all the fragments overlapping intronic and intergenic regions except those extending the 3' end of the gene, and spliced reads not supporting annotated exon-exon junctions, as they were considered to come from pre-mRNAs or unannotated transcripts. To avoid biases in the quantification of the isoforms due to reads mapping to internal priming (IP) locations in the genome, we discarded all the fragments that could arise from these sites. For that purpose, we scanned the whole genome using a custom Perl script and identified putative IP locations as regions containing six or more adenosines in a window of ten nucleotides. We discarded all the fragments located upstream from an IP site. In this case, we only considered IP sites located more than 60 nucleotides upstream of an annotated 3' end in the transcriptomic space.

3.2.3. Quantification of isoforms at the single-cell level

First, we used all genes with only one detected isoform to assess the empirical distance distribution of scRNA-seq reads with respect to annotated isoform 3' ends. Considering defined intervals of 30 nt, we calculated the distribution of read 3' ends relative to annotated 3' ends by dividing the number of distinct 3' ends in each bin in the transcriptomic space divided by the total number of 3' ends. Using this probability distribution, we assigned to each read a probability to come from a specific isoform based on its relative distance to the 3' end. Considering that each fragment is composed of one or more reads, we defined the probability of a fragment f to belong to an isoform k for a gene g as:

$$\mathbb{P}_f(k | g) = \omega_{k,g} \cdot \prod_{r \in R_f} \mathbb{P}_r(k | g)$$

where R_f is the set of reads associated with the fragment f , $\mathbb{P}_r(k | g)$ is the probability that read r is associated to the isoform k , and $\omega_{k,g}$ is the weight associated to the isoform k and gene g based on the pseudobulk quantification.

We computed the weight $\omega_{k,g}$ of an isoform k and gene g as

$$\omega_{k,g} = \frac{TPM_{k,g}}{\sum_{\kappa \in K_g} TPM_{\kappa,g}}$$

where K_g is the set of isoforms for the gene g and $TPM_{k,g}$ is the transcript per million counts of isoform k and gene g according to Salmon⁴⁴.

Hence, the probability of a set of fragments $F_{g,c}$ for gene g and cell c given a fixed set of isoforms K_g and associated relative abundances values (θ) is given by:

$$\mathbb{P}(F_{g,c} | K_g) = \prod_{f \in F_{g,c}} \sum_{\kappa \in K_g} \mathbb{P}_f(\kappa) \theta_{\kappa}^{g,c}$$

where $F_{g,c}$ is a set of fragments for a gene g and cell c , $\mathbb{P}_f(\kappa)$ represents the probability of a fragment f belongs to isoform κ , and $\theta_{\kappa}^{g,c}$ the isoform κ relative expression for a gene g and cell c .

Consequently, we estimated the isoform relative expression values (θ) by maximizing the log-likelihood function using a standard Expectation Maximization (EM) algorithm. The maximum likelihood estimators $\hat{\theta}$ for the isoform relative abundance values are given by:

$$\hat{\theta}^{g,c} = \underset{\theta^{g,c}}{\operatorname{argmax}} \sum_{f \in F_{g,c}} \log \left(\sum_{\kappa \in K_g} \mathbb{P}_f(\kappa) \theta_{\kappa}^{g,c} \right)$$

For each gene g and cell c , the EM algorithm proceeds as follows:

1. Initialization of value $\theta_k^{g,c}$ for each isoform k of the gene g as $\frac{1}{|K_g|}$
2. Iterate until convergence:
 - a. E-step: Calculation of posterior probability for each fragment $f \in F_{g,c}$ to belong to the isoform k given that it comes from a gene g and a cell c as

$$\mathbb{P}_f(k | g \cap c) = \frac{\mathbb{P}_f(k | g \cap c) \theta_k^{g,c}}{\sum_{\kappa \in K_g} \mathbb{P}_f(\kappa | g \cap c) \theta_{\kappa}^{g,c}}$$

- b. M-step: Estimate the isoform relative expressions $\theta_k^{g,c}$

$$\theta_k^{g,c} = \frac{\sum_{f \in F_{g,c}} \mathbb{P}_f(k | g \cap c)}{|F_{g,c}|}$$

The convergence state of the EM algorithm was settled by a stop criterion condition. This condition was reached when the maximum difference of the estimated relative abundance between two iterations was equal or lower than

0.001. All the isoforms with a null weight value were discarded from the annotation set. To generate isoform expression values for the iDGE per gene the estimated isoform probabilities were multiplied by the UMI counts assigned to the into the original DGE (Supplementary figure 2).

3.2.4. Isoform entropy and gene information content calculation

In order to deconvolute the DGE into an iDGE, SCALPEL estimates the isoform relative expression distributions for each gene and cell (θ , see above). Considering a fragment $\bar{f}$ across the set of all the fragments F , we can define the conditional probability of fragment $\bar{f}$ to originate from an isoform k given it comes from a gene g and cell c as:

$$\mathbb{P}_{\bar{f}}(k | g \cap c) := \frac{\sum_{\bar{f} \in F} \mathbb{P}_{\bar{f}}(k | g \cap c)}{|F|} = \theta_k^{g,c}$$

where $\theta_k^{g,c}$ is the isoform k relative expression estimated by SCALPEL for gene g in cell c . Taking this into account, the probability of fragment $\bar{f}$ to originate from isoform k of gene g given that it originates from cell c can be derived as

$$\mathbb{P}_{\bar{f}}(k | c) = \mathbb{P}_{\bar{f}}(k | g \cap c) \mathbb{P}_{\bar{f}}(g | c)$$

where the probability of fragment $\bar{f}$ to originate from any isoform of gene g given it originates from cell c can be calculated from the iDGE as:

$$\mathbb{P}_{\bar{f}}(g | c) = \frac{DGE_{g,c}}{\sum_{g \in G} DGE_{g,c}} = \frac{\sum_{\kappa \in K_g} iDGE_{\kappa,c}}{\sum_{\bar{\kappa} \in K} iDGE_{\bar{\kappa},c}}$$

Where $DGE_{g,c}$ denotes the number of UMIs assigned to gene g in cell c , G the set of all the genes, $iDGE_{\kappa,c}$ the number of UMIs assigned to isoform κ in cell c , K the set of all the isoforms across all the genes.

Applying Bayes' theorem, the probability of fragment $\bar{f}$ to originate from cell c given it originates from isoform k can be derived as:

$$P_{\bar{f}}(c|k) := \frac{P_{\bar{f}}(k | c) P_{\bar{f}}(c)}{P_{\bar{f}}(k)}$$

where the probabilities $P_{\bar{f}}(c)$ of fragment f to originate from cell c and $P_{\bar{f}}(k)$ of fragment f to originate from isoform k can be estimated from the iDGE corresponding column and row sum fractions, respectively.

$$\mathbb{P}_{\bar{f}}(c) := \frac{\sum_{g \in G} DGE_{g,c}}{\sum_{c \in C} \sum_{g \in G} DGE_{g,c}} = \frac{\sum_{k \in K} iDGE_{k,c}}{\sum_{c \in C} \sum_{k \in K} iDGE_{k,c}}$$

where C is the set of all the cells, and

$$\mathbb{P}_{\bar{f}}(k) := \frac{\sum_{c \in C} iDGE_{k,c}}{\sum_{k \in K} \sum_{c \in C} iDGE_{k,c}}$$

Given the cell-to-cell cluster mapping $\phi: C \mapsto L$, where L is the set of all cell clusters, the probability of fragment $\bar{f}$ to originate from a cell in cluster l , given it originates from isoform k can be derived by summing up the corresponding cell probabilities:

$$\mathbb{P}_{\bar{f}}(l | k) := \sum_{c \in C_l} \mathbb{P}_{\bar{f}}(c | k)$$

where C_l is the set of the cells associated to the cluster l .

Based on those conditional probabilities, each isoform's entropy across cell cluster is defined as:

$$H(k) := - \sum_{l \in L} \mathbb{P}_{\bar{f}}(l | k) \log_2(\mathbb{P}_{\bar{f}}(l | k))$$

If all fragments originating from a given isoform originate from cells of the same cell type, the corresponding entropy is minimal, at a value of

$$H_{\min} = 0$$

bits. If an isoform k is expressed perfectly equal across all clusters, the maximum isoform entropy

$$H_{\max} = \log_2 |L|$$

is reached, where $|L|$ denotes the number of cells clusters.

The isoform-level entropies were summarized to the gene level to quantify the randomness of a gene's isoform distribution across cell types:

$$\mathbb{H}(g) := \sum_{\kappa \in K_g} H(\kappa)$$

While the minimal gene entropy

$$\mathbb{H}_{\min} = 0$$

is gene-independent, its upper bound depends on the number of expressed isoforms $|K_g|$ and thus differs across genes:

$$\mathbb{H}_{\max}(g) = |K_g| H_{\max}$$

The difference from that maximum entropy defines a gene's information content across isoforms and cell types/clusters

$$I(g) := \mathbb{H}_{\max}(g) - \mathbb{H}(g)$$

3.2.5. Detection of differential isoform usage between cell clusters

For the identification of differential isoform expression between clusters of cells, we implemented the function *FindIsoforms*. This function aggregates the isoform counts using *AggregateExpression* from Seurat and retains only genes with two or more isoforms. All isoforms that represent less than 10% of the total gene expression in at least one of the conditions were discarded. Additionally, to reduce the number of false positives, we also discarded genes whose isoforms changed across conditions but had a very low expression variance (<5% standard deviation), which would otherwise be likely identified as significant DIU genes. For the remaining genes, the function performs a Chi-squared test to assess if the read distribution across isoforms in the same between clusters. We selected significant DIU genes with a false discovery rate (FDR) < 0.05.

3.2.6. Average 3' UTR length calculation

We extracted from the iDGE all protein-coding isoforms and computed the lengths of the corresponding 3' UTR regions from the reference annotation. For each gene g , we calculated the gene average 3' UTR length τ_g in a cell c as

$$\tau_{g,c} = \frac{\sum_{k \in K_g} \tau_k \text{ iDGE}_{k,c}}{|K_g|}$$

where τ_k is the 3' UTR length of isoform k , $iDGE_{k,c}$ are the UMI counts of isoform k .

Next, we calculated the average weighted 3' UTR isoform length within each cell τ_c for all the expressed isoforms as

$$\tau_c = \frac{\sum_{g \in \mathbb{G}} \tau_{g,c}}{|\mathbb{G}|}$$

where $\mathbb{G}$ is the set of all genes with protein-coding isoforms expressed.

3.2.7. Generation of read coverage plots

SCALPEL outputs a BAM file including the reads used for isoform quantification. Within SCALPEL framework, we have implemented the function *CoveragePlot* in R to visualize the read coverage on the isoforms using a transcriptome annotation in GTF format and BAM files. We generated the visualization tracks using the R Gviz library v1.46.0¹⁷⁶.

3.2.8. Generation of synthetic data for benchmark analysis

We used Splatter¹⁷⁷ to simulate 12320 isoforms across 6000 cells representing two equally probable cell populations (group.prob = 0.5) using the function *splatSimulatedGroups*. For these cells, we generated 3 datasets with distinct sequencing depth: low (lib.loc: 29,000 counts, lib.scale: 0.5), medium (lib.loc: 39,000 counts, lib.scale: 0.5), high (lib.loc: 120,000 counts, lib.scale: 0.5). Dropout effects were modeled using the experimental dropout function with the parameters *shape* and *midpoint* adjusted to reflect 10x genomic profiles in the 10x scRNA-seq mouse dataset (shape: 0.5; midpoint: 0.4). For each simulation,

genes were stratified into five regulatory categories based on their number of isoforms and their expression dynamics: unique isoform with no change, unique isoform with expression change, multi-isoform with no change, multi-isoform with coordinated expression change, and multi-isoform with isoform-specific regulation (DIU) (Supplementary table 2 and Figure 17b). Furthermore, genes were sampled across five expression quantiles to ensure even coverage across expression levels.

To complement the synthetic iDGE matrices, we generated synthetic FASTQ files that reflect realistic read structure and position distribution observed in the mouse 10x scRNA-seq data. Considering the mapped data, we extracted the genes, isoforms, and fragments, as well as the distance to the transcript's 3' end for all reads assigned to at least one transcript. Next, we used a custom bash script to sort these data and extract the following three empirical distributions: i) the number of reads per fragment; ii) their distance to the fragment's 3' end; and iii) the distance to the transcripts 3' end for all fragments that were unambiguously mapped to a gene expressing a single isoform. From these distributions and the counts from the corresponding synthetic iDGE, fragments and reads were sampled using a custom R script. Then, a second custom awk script was used to extract the corresponding synthetic read sequenced from a FASTA file providing the transcripts' spliced sequences. This single FASTA file was finally converted to a pair of FASTQ files adding '@' quality scores and transforming the placeholder cell and fragment indices to valid 10x barcodes using a custom Python script. All custom scripts used to generate the synthetic FASTQ files are implemented within the scr⁴eam pipeline (see code availability).

3.2.9. Analysis of mouse spermatogenesis 10x scRNA-seq data

We downloaded the 10x scRNA-seq samples from male mouse germline¹⁷⁸ from GEO database (accession number GSE104556) and processed them using Cell Ranger v7.1.0⁹⁷, using mm10 mouse genome assembly¹⁷⁹ and GENCODE M21¹⁷¹ as reference annotation. We merged the processed data and analyzed them jointly using Seurat v5.0.0¹⁸⁰. We restricted the analysis to the cells with a number of UMIs counts superior to 500, inferior to 80,000 and less than 5% of mitochondrial genes. The final Seurat object contained 2,174 cells and 22,158 genes. We performed dimensionality reduction on the DGE using 2,000 genes as variable features and used the first 9 principal components (PCs) to build the kNN graph and compute a UMAP. We used the function *FindClusters* with a resolution of 0.02 to identify 3 clusters. We compared our clustering results with the cell clustering from a previous analysis on the same dataset by Schulman et al³⁸ and annotated the defined cell clusters in our analysis as ES, RS and SC according to their previous corresponding annotation. Next, we filtered the input BAM file to retain the 2,174 filtered cells using *samtools view -D CB* from samtools v1.19.2¹⁷³ and extracted the DGE matrix for SCALPEL analysis. Following the execution of SCALPEL, we generated a Seurat object containing the same 2,042 cells, 23,151 genes and 76,419 isoforms. For downstream analyses, we discarded all isoforms expressed in less than four cells and cells with less than three isoforms expressed using *CreateSeuratObject* function (option min.cells=4, min.features=3). The final Seurat object included 51,767 isoforms of 17,525 genes. We performed dimensionality reduction on the iDGE using 2,000 isoforms as variable features, and we used 9 PCs for the generation of the kNN graph for

the UMAP. We used a resolution of 0.02 to perform the clustering and identified three cell populations. We used the approach described above to annotate the cell clusters. We compared the iDGE clusters with the DGE clusters using a Jaccard index.

3.2.10. Differentiation and characterization of iPSCs and NPCs using Drop-seq technology

We differentiated human iPSCs towards NPCs using a protocol previously established in the lab¹⁵⁸. Briefly, we differentiated iPSCs to neuroepithelial cells over a period of ten days by dual SMAD inhibition using neural maintenance medium (1:1 ratio of DMEM/F-12 GlutaMAX (Gibco, #10565018) and neurobasal (Gibco, #21103049) medium complemented with 0.5x N-2 (Gibco, #17502048), 0.5x B-27 (Gibco, #17504044), 2.5 µg/ml insulin (Sigma, #I9278), 100 mM L-glutamine (Gibco, #35050061), 50 µM non-essential amino acids (Lonza, #BE13-114E), 50 µM 2-mercaptoethanol (Gibco, #31350010), 50 U/ml penicillin and 50 mg/ml streptomycin (Gibco, #15140122)) supplemented with 500 ng/ml noggin (R&D Systems, # 3344-NG-050), 1 µM Dorsomorphin (StemCell technologies, #72102) and 10 µM SB431542 (Calbiochem, # 616461). After the initial neural induction step, we differentiated the cells to NPCs by neural maintenance medium replacement up to day 22. We cryopreserved iPSCs and NPCs in fetal bovine serum or neural maintenance medium supplemented with 10% of DMSO for iPSC or NPCs respectively. Before single-cell encapsulation using NADIA instrument (Dolomite Bio), we thawed, filtered and counted the samples. We retrotranscribed RNAs captured with the oligo(dT) for cDNA library preparation. Finally, we sequenced the final Illumina tagged libraries on an Illumina NextSeq

550 sequencer using the NextSeq 550 High Output v2 Kit (75 cycles) (Illumina, #20024906) in paired-end mode; read 1 of 20 bp with custom primer Read1CustSeqB⁹⁵, read 2 of 64 bp and 8 bp for i7 index.

3.2.11. scRNA-seq analysis of Drop-seq data

We processed the iPSC and NPC scRNA-seq libraries using Drop-seq tools v2.5.1⁹⁵ pipeline to generate DGE matrices. We merged the FASTQ files containing paired end reads into a single unaligned BAM file using Picard tools v2.27.4¹⁶¹. We tagged the reads with cell and the molecular barcodes, trimmed them at the 5' end to remove adapter sequences and at the 3' end to remove poly(A) tails, and mapped them to the human genome (*GRCh38*) with STAR v2.7.10¹⁸¹. We tagged the resulting BAM files with the annotation metadata using the human GENCODE v41¹⁸² annotation as reference. Finally, we performed the cell barcode correction using the programs *DetectBeadSubstitutionError* and *DetectBeadSynthesisErrors* with default parameters. To estimate the number of cells obtained, we used a knee plot considering the top 3,000 cell barcodes and generated a DGE count matrix for each sample.

We used Seurat v5.0.0¹⁸⁰ and R 4.3.2¹⁸³ to merge the DGEs and preprocess the scRNA-seq data. We discarded all genes expressed in less than four cells and all cells with less than three genes expressed. We also discarded low quality cells with less than 300 UMIs, less than 300 genes and more than 5% mitochondrial gene, and cell artifacts with more than 20,000 UMIs and 7,000 genes. The final Seurat object contained 2,535 cells and 19,103 genes. We performed a dimensionality reduction analysis using the 2,000 most variable genes to calculate 50 PCs. We used the *ElbowPlot* function to manually inspect the

amount of variability explained by each PC and selected the first 9 PCs to build the kNN graph and compute the UMAP plot.

3.2.12. Benchmark analysis

We downloaded each of the benchmarked tools from their respective GitHub repository. For each tool, all commands for its default execution were integrated into Nextflow workflows and were executed using the default parameters indicated by the authors. We performed the benchmark analysis on the preprocessed mouse spermatogenesis scRNA-seq¹⁷⁸ data using the GENCODE vM21¹⁸² GTF and transcriptome FASTA files as reference annotation. Additionally, we performed a second benchmark analysis on the preprocessed neuronal cell differentiation scRNA-seq Drop-seq data using the analogous reference files obtained from GENCODE v41¹⁸². As scAPA only considers disjoint 3' UTR regions annotation from the hg19 version of the human genome for the annotation of its quantified peaks, we generated a new annotation of non-overlapping 3' UTR regions using GENCODE v41 and *GenomicRanges* R package. We intersected all the peaks detected by scAPA³⁸ following its peak calling process to this new reference annotation. Following Dapars2 v2.1¹⁸⁴ default procedure⁸⁰, we downloaded the gene region annotation reference for the human and mouse genome (GRCh38, mm10) using the UCSC Table browser. Then, we extracted 3' UTR regions from the gene annotation using Dapars2 script *DaPars_Extract_Anno*. Finally, we calculated the raw percentage of distal PAS sites usage index (PDUI) values using DaPars2 script *DaPars2_Multi_Sample_Multi_Chr* and provided them to scDaPars⁸⁰ to infer their expression at single-cell level. scUTRquant⁷⁹ was executed using the target

transcriptome annotation files provided in the GitHub repository for the human and mouse genomes (hg38 and mm10). These files included high-confidence cleavage sites called from the Human Cell Landscape and Mouse Cell Atlas dataset. Additionally, we reran scUTRquant⁷⁹ using a custom target transcriptomic annotation generated from the input genome annotation using the Bioconductor package *txcutr* (v1.8.0)⁷⁹. For each tool, we performed a differential peak or isoform usage analysis with default parameters and an adjusted p-value threshold of 0.05 for all genes with at least two peaks or isoforms detected. For each comparison test, we generated the UpSet plots using the R library UpSetR v1.4.0¹⁸⁵ for the set of DIU genes co-detected by at least two tools. We applied a cutoff threshold of 10 genes for each intersection set.

3.2.13. Isoform validation using nested PCRs

To validate the changes of isoform usage between the iPSCs and NPCs predicted by SCALPEL, we performed nested PCR as previously described^{186,187}. One µg of total mRNA extracted using Maxwell® RSC simplyRNA Cells kit protocol (Promega Corporation, #AS1340) was used as input RNA for the cDNA synthesis using an oligo(dT)-adaptor sequence TAP-VN as a primer for the reverse transcription. For the first nested PCR, one µL of 1:10 cDNA dilution was used, with a gene-specific primer (GSP) which is shared by all isoforms and an adaptor primer (AP) as a reverse primer. The second nested PCR was performed with one µL of a 1:10 dilution of the first PCR using a second gene-specific primer and a second adaptor primer (MAP) as a reverse primer. This second nested PCR reaction that anneals 3' to the first GSP is essential to reduce the amplification of undesired products¹⁸⁸. The resulting nested PCR products are

resolved by an agarose gel. Primer sequences are provided in Supplementary Table 21.

3.2.14. Identification of miRNA signatures in differentially regulated isoforms

To obtain isoform-level identification of miRNA target sites, we overlapped the genome-wide miRNA target site annotation included in the MBS database¹⁸⁹ with the GENCODE annotation reference v41¹⁸² (hg38). Target sites of different miRNAs were grouped by their seed sequences according to miRBase v22.1¹⁹⁰. The seed-target isoform pairs were used for downstream analyses. Using a set of neurogenesis-related miRNA¹⁹¹, we filtered genes displaying changes in isoform usage between iPSCs and NPCs as predicted by SCALPEL with at least one isoform targeted by a neurogenesis-related miRNA and one non-targeted isoform. We normalized the isoform abundances by the number of cells in NPCs and iPSCs and used these values to calculate their log2 fold changes (log2FC). For each miRNA, we used a Kolmogorov-Smirnov test to check for differences in the cumulative distribution of log2FC between targeted and non-targeted isoforms from the same set of genes. The resulting p-values were FDR-adjusted using Benjamini-Hochberg correction.

3.2.15. Identification of novel cell populations in the mouse 10x dataset

We increased the clustering resolution in the gene and the isoform-based analyses to identify new cell states. We tested different number of variable features in the DGE and iDGE datasets and assessed the resulting clustering

using a range of resolutions. For each parameter set, we evaluated the concordance between the DGE and iDGE clustering by calculating the Jaccard index between the cell barcodes in each cluster. Based on this analysis, we selected 4,000 variable features and a clustering resolution of 0.5 for the DGE, and 6,000 variable features with a resolution of 0.9 for the iDGE. Using the markers from Lukassen¹⁹², we annotated the cell clusters of the gene-based analysis. For the isoform analysis, we calculated a Jaccard index score between all gene and isoform clusters (Supplementary figure 3a) and assigned to each isoform-based cluster the identity of the most similar cluster. We performed differential isoform analysis using *FindAllMarkers* function (option `min.pct=3`, `p_val_adj<0.05`) on the iDGE to identify isoforms markers for each cluster. We visualized the top 100 isoform markers average expression within each cluster in the iDGE using the *Heatmap* function from the *ComplexHeatmap*¹⁹³ package along the average expression of their corresponding genes in the DGE clusters (Supplementary figure 3d).

3.2.16. Differential gene expression analysis between RS populations

We performed a differential gene expression analysis between the RS6 and other RS populations using a pseudo-bulk approach. We aggregated gene expression values across cell clusters using the Seurat function *AggregateExpression* to generate a pseudo-bulk count matrix. After removing genes with zero counts, we defined a design matrix indicating the condition for each sample (RS6 - RS cells). Using DESeq2 R library⁷², we constructed a DESeq2 object with *DESeqDataSetMatrix* function and performed differential expression analysis

with DESeq *function*. We identified significantly differentially expressed genes with an adjusted p-value inferior to 0.05 and an absolute log2 fold-change value superior to 0.5.

3.2.17. GO term enrichment analysis

We performed GO term enrichment analysis using the R package *enrichR* v3.2¹⁹⁴. The reference used for the enrichment analysis was the *GO_Biological_Process_2023*¹⁹⁵. We selected all the GO terms with an adjusted p-value inferior to 0.05 and visualized the associated adjusted p-value and *enrichR* combined scores using *ggplot2*¹⁹⁶.

3.2.18. Trajectory inference analysis

We used the iDGE-based Seurat object to derive a pseudo temporal ordering of the cells using Monocle3 v1.3.4⁹². First, we converted the Seurat object into a *CellDataSet* object which includes the cluster annotation and the UMAP embedding previously computed. Next, we fit the principal trajectory graph within each cluster partition using the function *learn_graph*. Finally, we calculated the pseudotime values using the function *pseudotime* and the ES cells as root cells.

3.2.19. Analysis of paired short and long scRNA-seq datasets

We downloaded the FASTQ files for Illumina short-read and PacBio long-read datasets of four samples from the P28 developmental stagepoint from Joglekar et al.⁷¹ from the Knowledge Brain Map database (RRID:SCR_016152). These datasets correspond to two samples from the hippocampus (M1-HIPP, M2-HIPP) and two samples from the visual cortex (M1-VIS, M2-VIS).

We processed the Illumina short reads FASTQ files using Cell Ranger v7.1.0⁹⁷ with the mm10 mouse genome assembly¹⁷⁹ and generated Gene x Cell count matrices. We merged the hippocampus and visual datasets and analyzed them jointly using Seurat v5.0.0¹⁸⁰. We filtered out cells with less than 1,000 and more than 25,000 UMI counts and cells with more than 15% mitochondrial genes content. After quality control, we retained 43,094 cells for 32,285 genes. We performed dimensionality reduction on the DGE using 2,000 genes as variable features and integrated the data using Harmony¹³¹ to correct for sample-specific batch effects⁷⁸. Next, we used the first 30 principal components (PCs) to build the kNN graph and compute a UMAP. We applied the Seurat function *FindClusters* with a resolution of 0.05 to identify 14 clusters and annotated them using known marker genes from the CellMarkers database¹⁹⁷. Then, we reduced the dataset by selecting the set of barcodes corresponding to Excitatory neurons, Astrocytes, Neural progenitor cells (NPCs), Microglia and Oligodendrocytes to retain 32,356 cells. We repeated the dimensionality reduction analysis using 2000 variable genes and 7 PCs and applied a clustering resolution of 0.03 to generate 5 distinct clusters.

We processed the PacBio long-read FASTQ files using the scisorseqr package v5.0.1⁷¹. First, we deconvoluted the samples with the 32,356 cell barcodes from the short read single-cell analysis using the function *GetBarcodes*. Then, we aligned the filtered reads to the mm10 mouse reference genome using the function *MMalign*, to generate a BAM file for each cell type. We used function *MapAndFilter* to identify full-length, spliced reads supported by both CAGE and poly(A) peaks using annotation references from FANTOM5¹⁹⁸ and poly(A)Site¹⁹⁹. We set the *filterFullLength* parameter to TRUE to retain only 5'- and 3'-complete

reads. Finally, we quantified isoform usage for cell types using the *IsoQuant* function (Iso=TRUE, TSS=TRUE, PolyA=TRUE).

We applied SCALPEL on the filtered 10x short-read single-cell dataset, using mouse GENCODE M24 annotation²⁰⁰ reference to generate an isoform-based count matrix (iDGE). We filtered all the isoforms expressed in less than four cells to retain 43,140 isoforms across 32,356 cells. Using Seurat¹⁸⁰, we performed a dimensionality reduction analysis by selecting 9500 variable features and 10 principal components. We integrated the samples with Harmony¹³¹ with default parameters to correct for batch effects, then clustered the iDGE data with a resolution of 0.05. We identified gene and isoform markers for each cell-type within the data using the Seurat function *FindMarkers* with an adjusted p-value of 0.05, and avg.log2 fold-change cutoff of 0.1 (Supplementary table 6, Supplementary table 7). Finally using SCALPEL function *FindIsoforms* with default parameters (Adjusted p-value:0.05, threshold.var:0.05, threshold.abund: 0.1), we identified DIU genes between the distinct cell-types (Supplementary table 8).

3.2.20. Statistical analyses

We performed all the statistical tests in this manuscript using R v4.3.2¹⁸³.

3.2.21. Data availability

The Drop-seq data from iPSCs and NPCs generated within this project are available on GEO under accession number GSE268222.

3.2.22. Code availability

SCALPEL is implemented in Python and R. The code of SCALPEL and the pipeline to benchmark all tools presented in this manuscript are available on GitHub (<https://github.com/plasslab/SCALPEL>).

4. Results

4.1. Objective 1: Optimize single-cell transcriptomics protocols to work with iPSC-derived neuron and glial cells

All experiments and analyses described under Objective 1 have been published in the open-access article by Gutiérrez-Franco*, Ake*. et al. (2023)ⁱ, available in *Communications Biology* ²⁰¹.

4.1.1. Impact of sample preservation methods on single-cell library quality and cell composition

4.1.1.1. Preservation methods influence the complexity of captured transcriptomes

To evaluate the impact of cell preservation on single-cell RNA-seq quality, we generated single-cell suspensions from hiPSC-derived neural progenitor cells to neuroprogenitors (NPCs) and processed them under six different conditions of preservation/fixation methods: fresh (non-preserved), DMSO cryopreserved, methanol-fixed, ACME-fixed, vivoPHIX-fixed, and CellCover-treated. For each condition and sample, we used the same library preparation protocol and encapsulation and with the NADIA equipment, a commercial Drop-seq setup. The obtained datasets were then evaluated using different metrics to assess their quality. Inspection of the cDNA profiles showed that ACME and vivoPHIX samples had less cDNA and smaller fragments than the rest of the libraries (Figure 11a), which is consistent with RNA degradation. Next, we evaluated the impact of the preservation methods on library complexity. We used samtools¹⁷³ to downsample

ⁱ Gutiérrez-Franco and Ake contributed equally to this work.

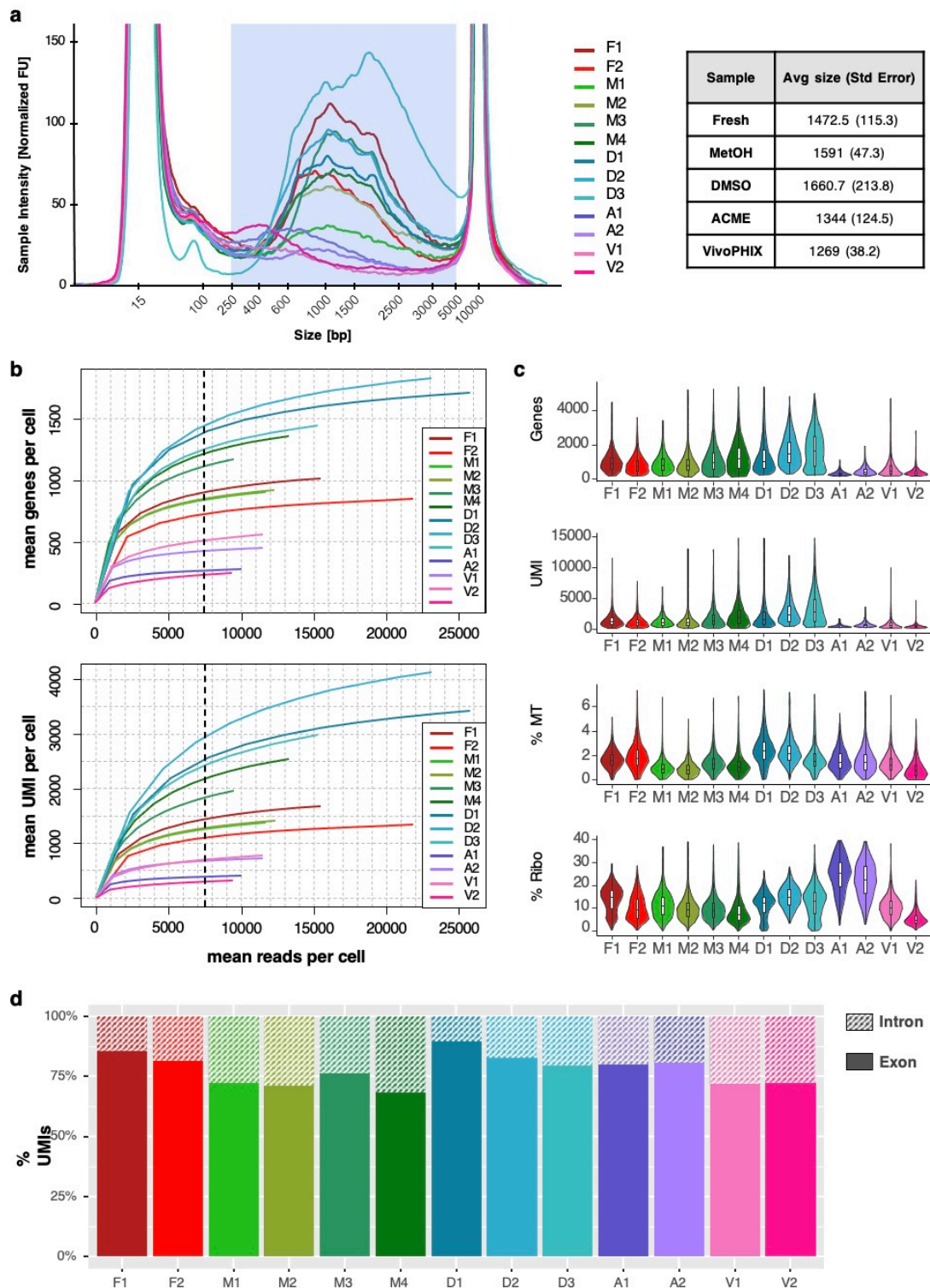


Figure 11. Systematic comparison of the effects of cellular preservation on library quality **a.** Trace of the libraries obtained after encapsulation, including fresh (F1 and F2), methanol (M1, M2, M3, and M4), DMSO (D1, D2, and D3), ACME (A1 and A2), and vivoPHIX (V1 and V2) samples. The libraries of the samples fixed with vivoPHIX and ACME have smaller fragment sizes and less RNA, which is consistent

with RNA degradation. **b.** Downsampling plots showing the number of genes and UMIs as a function of sequencing depth (mean reads per cell). In both cases, DMSO (blue) and methanol (green) libraries have a depth equivalent or higher than that of fresh samples (red). **c.** Violin plots showing the number of genes, UMIs, the percentage of mitochondrial content (% MT), and ribosomal content (% Ribo) of the cells for each sample after discarding low-quality cells and doublets. The boxplots included inside the violin plots summarize the data distribution. The upper and lower sides of the box represent the 1st and 3rd quartiles. The line in the middle corresponds to the median. Lines extend no further than 1.5 the interquartile range. **d.** Barplots showing the percentage of intronic and exonic UMIs assigned to cells for each of the libraries. The number of intronic reads range from ~10% in DMSO sample D1 to ~30% in methanol sample M4. High intronic UMI fractions are indicative of cellular RNA leakage or nuclear RNA enrichment.

each of the unaligned BAM files to produce files containing 10%, 20%, 30%, etc. of the original dataset. Each of these subsamples was then processed using the same computational pipeline to generate downsampled digital gene expression (DGE) matrices. Our results show that at the same sequencing depth, DMSO cryopreserved cells and methanol-fixed cells yield a comparable or higher number of genes and unique molecular identifiers (UMIs) per cell than fresh cells. In comparison, at 7500 reads per cell, ACME and vivoPHIX samples have around 40% of the Genes and UMIs obtained in fresh cells (Figure 11b). The higher number of genes and UMIs in the DMSO samples D1 and D2 and the methanol-fixed samples M3 and M4 is due to differences in the capture efficiency of the beads used for the encapsulation and not to an intrinsic higher capture efficiency in these experimental conditions (Supplementary figure 5).

The lower complexity of vivoPHIX and ACME libraries is also reflected in the proportion of low-quality cells in the samples, which have few UMIs and genes

detected per cell (Figure 11c, Table 1), and may correspond to empty droplets or droplets containing broken cells²⁰². While the number of low-quality cells

Sample	Cells	Filtered cells	% cells kept	Mean genes per cell	Mean UMI per cell	% MT genes	% Ribo genes
F1	1764	1611	91.33	1043	1738	1.65	14.1
F2	1327	1142	86.06	889	1390	1.90	10.1
M1	1841	1613	87.62	970	1479	0.961	11.3
M2	1752	1568	89.50	954	1469	0.886	9.60
M3	1833	1697	92.58	1185	1977	1.39	9.71
M4	1305	1172	89.81	1364	2532	1.16	8.61
D1	1234	1139	92.30	1692	3271	2.31	14.8
D2	1342	1228	91.51	1787	3943	1.78	12.4
D3	1564	1405	89.83	1352	2577	2.49	11.4
A1	1313	729	55.52	373	545	1.61	24.5
A2	1524	1121	73.56	535	886	1.55	22.4
V1	1573	1122	71.33	669	937	1.31	10.6
V2	3053	1323	43.33	416	529	0.877	5.51

Table 1: Cell quality statistics.

discarded is around 10% of cells in fresh, DMSO, and methanol-fixed cells, this number increases up to 29% and 49% in ACME and vivoPHIX samples, respectively (Table 1, Supplementary table 1). After discarding low-quality cells, the remaining vivoPHIX and ACME cells still show clear differences in quality (Figure 11c and Table 1). Interestingly, the percentage of UMIs mapped to mitochondrial genes was similarly low across all samples, suggesting that the lower number of captured RNAs in these samples may not be due to RNA leakage or cellular damage²⁰² but rather be related with RNA capture efficiency

or RNA degradation. In addition, ACME samples showed a much higher fraction of UMIs mapped to ribosomal proteins than any of the other samples (Figure 11c), which has also been previously associated to low-quality cells or technical artifacts²⁰³.

One of the reasons why vivoPHIX and ACME fixed cells have less RNAs per cell could be that the fixation method facilitates cell breakage. In this case, we would expect lower library complexity and a higher fraction of reads coming from intronic regions, which mainly come from pre-mRNAs and are enriched in nuclei. In our samples, the fraction of intronic reads coming from each sample was variable across preservation methods, ranging from 10% in DMSO samples to ~30% in methanol and vivoPHIX (Figure 11d) samples. The higher fraction of intronic reads in vivoPHIX and methanol samples is indicative of nuclear RNA enrichment^{102,204}. Yet, considering that the number of genes and UMIs detected in methanol is, on average, around three times higher than in vivoPHIX samples (Supplementary table 1), RNA leakage is likely not the cause of this bias.

Together, these results demonstrate that vivoPHIX and ACME preservation methods significantly affect the quality of the single-cell transcriptomes obtained from human NPC populations although this cannot be explained by an increased RNA leakage or cell breakage.

4.1.1.2. DMSO cryopreservation alters the cell composition of hiPSC-derived cell populations

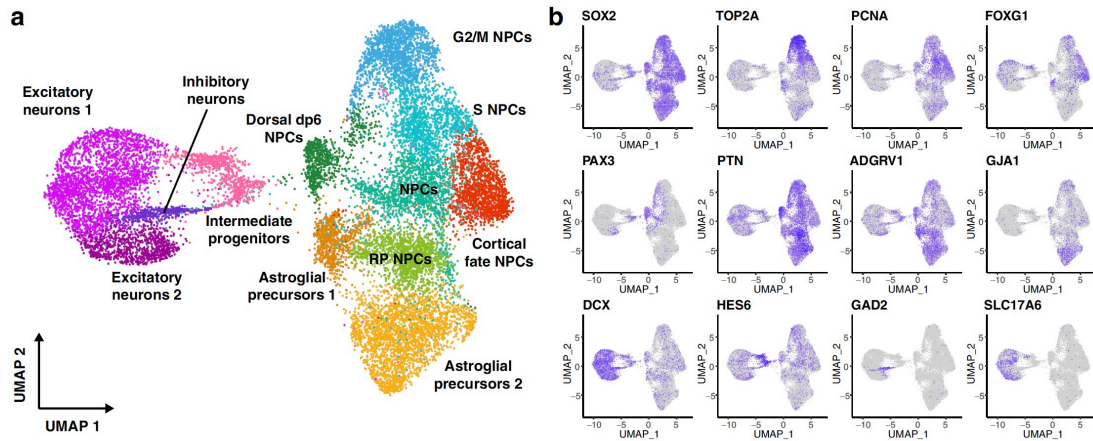


Figure 12. scRNA-seq identifies diverse progenitor and immature neuron populations. **a.** UMAP plot showing the cell populations identified in the NPC samples, which include different types of NPC populations, Astroglial precursors, intermediate progenitors, and both excitatory and inhibitory immature neurons. **b.** Feature plots of known marker genes that have been used to identify the cell populations in (a) including NPC markers (SOX2), proliferation markers (TOP2A and PCNA), cortical fate markers (FOXG1), dorsal fate markers (PAX3), roof plate markers (PTN), Astroglial markers (ADGRV1 and GJA1), immature neuronal markers (DCX), intermediate progenitor markers (HES6), inhibitory (GAD2), and excitatory (SLC17A6) neuronal markers.

After assessing the overall quality metrics of the samples, we investigated whether preservation methods affect the cellular composition of the samples. For that purpose, we pooled all the samples and analyzed them together. Initial analysis showed a strong batch effect driven mainly by the preservation method used. As can be seen in Supplementary table 4, before integration, the cells from different experiments occupy different regions of the uniform manifold approximation and projection (UMAP) plot. Thus, we used Harmony¹³¹ to integrate the datasets and identify the cell populations obtained from the hiPSC

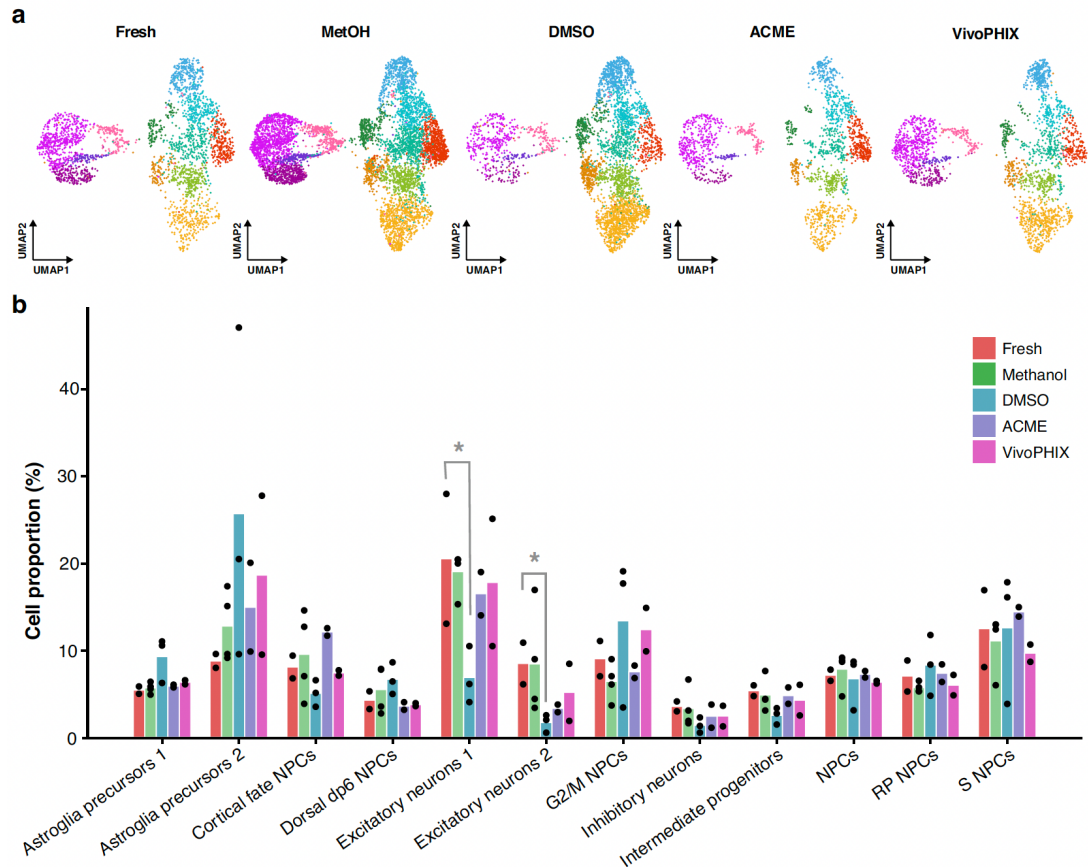


Figure 13. DMSO cryopreservation impacts the cell composition of scRNA-seq datasets **a.** UMAP plots showing the distribution of cells across clusters for each of the fixation methods. Clusters are colored as in Fig. 3a. DMSO-preserved samples show a strong depletion of cells in neuronal clusters compared to all other samples. **b.** Barplot showing the average proportion of cells assigned to each cluster for each method. The height of the bars represents the average for all samples preserved using the same method. Black dots represent the cell proportion in individual samples. * Indicates samples with a significant difference in cell proportion in a particular sample relative to fresh samples as predicted by scCODA, which means that the average score of the sample is below zero according to the scCODA model considering a false discovery rate <0.05.

differentiation (Figure 12a). After batch correction, we identified 12 cell populations corresponding to proliferating progenitors, NPCs, Astroglial precursors, intermediate progenitors and different types of neurons characterized by the expression of specific marker genes (Figure 12b, Supplementary Figure

4, Supplementary Figure 7). All the clusters identified were present in all the individual samples. Yet, the relative abundance of each of the cell populations changed depending on the preservation method used (Figure 13a, Supplementary figure 7).

To investigate if these changes were due to experimental biases or they were systematic biases due to the fixation method, we performed a compositional analysis of the samples with scCODA¹⁶⁵, a recently developed tool that can reliably identify changes in single-cell datasets even with a low number of replicates. In contrast to other methods, scCODA models compositional bias of the sample as a whole and not for each cluster independently. This prevents wrongly identifying cell proportion changes due to the depletion of a single-cell population, which would artificially result in the increase of all other cell populations in the sample (Figure 13b). To identify compositional changes, we chose as reference group fresh samples, so that the results will indicate if we find compositional changes relative to this group. The results from this analysis highlight a significant depletion of excitatory neurons in DMSO cryopreserved samples compared to fresh samples, while we do not find significant compositional biases in the samples preserved using any of the other methods (Figure 13b).

4.1.2. Preservation methods induce biases in the gene expression profile of cells

Finally, we investigated if preservation methods induced changes in gene expression that could affect the comparison among samples. A comparison of gene expression across samples shows high correlation among all samples (Pearson correlation coefficient $R \geq 0.8$), although ACME and vivoPHIX samples have slightly lower correlations with all other samples (Supplementary figure 8). This lower correlation can be explained by global or cell-type-specific changes in gene expression but also by compositional biases. To investigate this further, we generated pseudobulk counts for each of the clusters for each sample separately and performed a correlation analysis at the cluster level. Our analyses show that the fixation method induces biases in the clustering of cell populations across samples (Supplementary Figure 9).

However, the high similarity between different clusters, i.e., NPC populations, makes cell clusters preserved with a particular method cluster together. To address this issue, we evaluated the clustering of samples for each cell cluster independently using sigclust2¹⁷⁰, a statistical method designed to test the statistical significance of hierarchical clustering. As can be seen in Figure 14, in all cases methanol samples clustered with fresh samples. In contrast, in 8 of the 12 cell populations identified, several vivoPHIX and ACME samples cluster separately from fresh samples. This analysis thus confirms that the overall expression profile of methanol-fixed cells in each cluster is more similar to that of fresh cells than that of cells preserved using other methods.

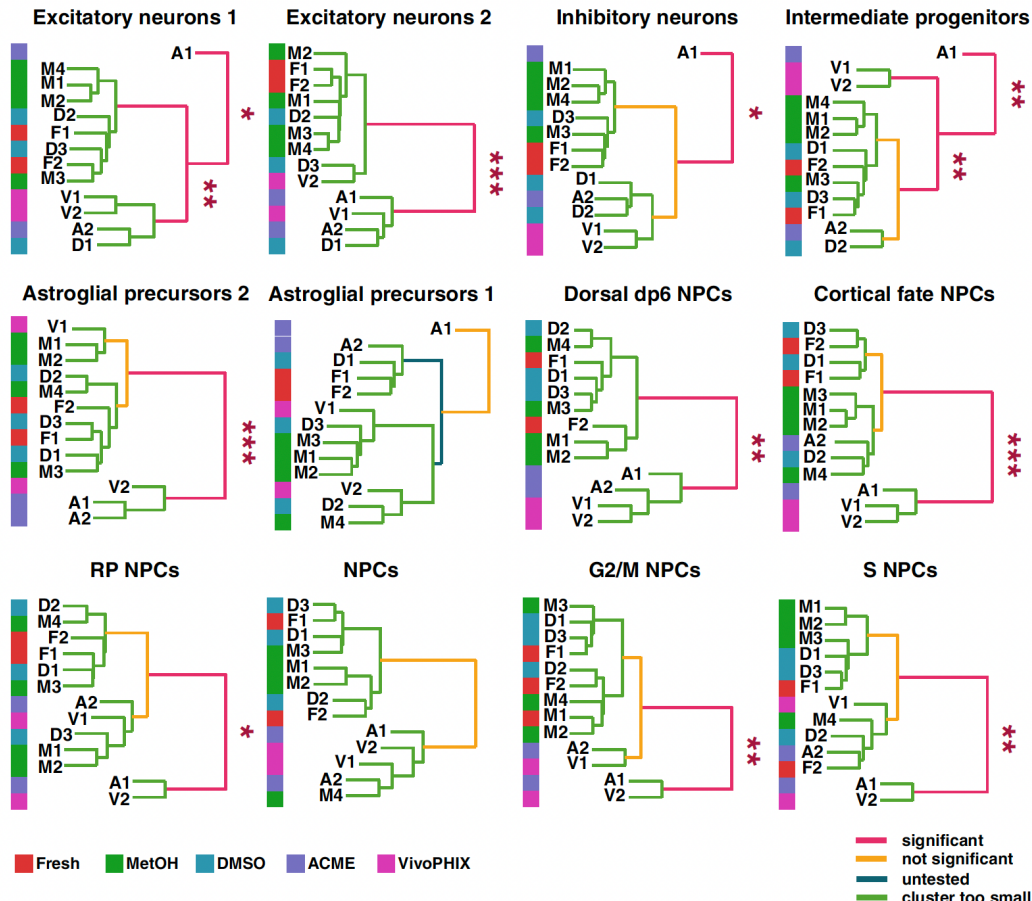


Figure 14. Computational analysis identifies fixation-related biases in cell clustering. Dendrogram showing the similarity of the gene expression profiles of the cells from the different samples belonging to the same cell cluster. The samples are labeled as follows: fresh (F1 and F2, red color), methanol (M1, M2, M3, and M4, green color), DMSO (D1, D2, and D3, blue color), ACME (A1 and A2, purple color) and vivoPHIX (V1 and V2, magenta color). The significance in the hierarchical clustering is assessed using the shc test implemented in sigclust2. Significant branches in the dendrogram are highlighted in pink. *, **, and *** mark the branches with a p-value smaller than 0.05, 0.01, or 0.001, respectively. Not-significant branches are highlighted in yellow and not-tested branches in blue and green.

Previous studies have shown that dissociation and preservation methods can induce cellular stress that is reflected at the transcriptomic level^{135,139}. Accordingly, we checked the expression of Immediate Early genes (IEGs) and apoptosis markers in our datasets. The apoptosis gene signature was higher in

ACME, vivoPHIX, and DMSO samples compared to fresh samples, and higher in DMSO than methanol-fixed samples, although these differences were minimal (Figure 15a). All fixed samples had higher expression of IEGs than fresh samples, although DMSO cryopreserved cells showed higher expression of IEGs than all the other samples (Figure 15b). This result indicates that freezing and thawing processes stresses the cells in a way that is globally reflected on their transcriptomic profile.

To investigate if cell preservation induced additional expression biases at the individual cluster level, we used muscat¹⁶⁶ to identify cell-type-specific differentially expressed genes (DEGs) in fixed samples compared to fresh (Supplementary table 3). Our analysis shows that the number of differentially expressed genes (DEGs) is very different across fixation methods. Whereas vivoPHIX samples present many DEGs in all clusters, the amount of significant DEGs is close to zero in DMSO samples (Figure 15c, Supplementary table 3). We used gene ontology term enrichment analysis to investigate if different fixation methods would introduce biases in gene expression associated to particular biological functions. Our results did not find any significant terms overrepresented in genes consistently up or downregulated across multiple cell clusters, suggesting that the effect that the fixation methods have on gene expression across clusters is not linked to particular cell functions or locations.

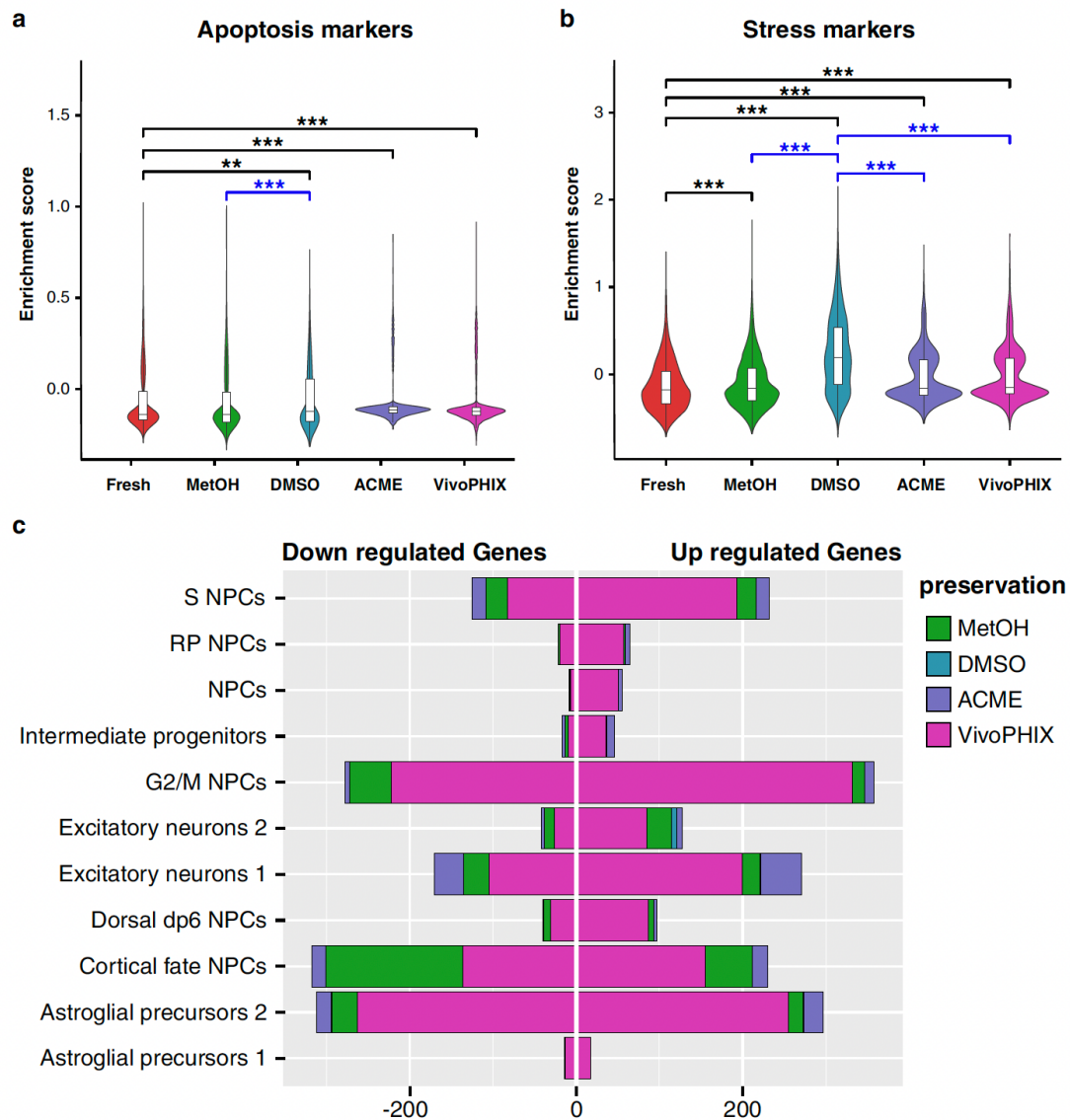


Figure 15. Fixation/preservation methods induce sample-specific expression biases. **a, b.** Violin plots showing the distribution of the enrichment scores of apoptosis (**a**) and stress (**b**) signatures across sample preparation methods. The boxplots included inside the violin plots summarize the data distribution. Upper and lower sides of the box represent the 1st and 3rd quartiles. The line in the middle corresponds to the median. Lines extend no further than 1.5 the interquartile range. **a.** The apoptosis enrichment score is higher in DMSO, ACME, and vivoPHIX samples compared to Fresh samples (black asterisks), and higher in DMSO compared to methanol (blue asterisks). **b.** The stress signature is higher for all fixation/preservation methods compared to fresh samples (black asterisks), and also higher in DMSO compared to all other fixation methods (blue asterisks). In all cases, statistical significance was tested using a one-tailed Wilcoxon rank-sum test. Comparisons

against fresh are marked with black * and comparisons of DMSO against other fixation methods are in blue. ** indicates a p-value <0.01 ; *** indicates a p-value <0.001 . **c.** Barplot showing the number of significantly upregulated and downregulated genes for each fixation method in each cell cluster with a $\log_2\text{fc} \geq |0.58|$ and an adjusted Wald test P-value <0.05 . vivoPHIX and methanol samples have more DEGs across

4.2. Objective 2: Implementation of a computational method for characterization of APA with scRNA-seq datasets

All experiments and analyses described under Objective 2 have been published by Ake*. et al. (2025)ⁱⁱ in Nature Communication²⁰⁵.

4.2.1. SCALPEL: A pipeline for isoform quantification at single-cell resolution

SCALPEL is a new computational Nextflow workflow²⁰⁶ to quantify transcript isoforms from single-cell data. It takes as input the digital gene expression matrix (DGE) generated by a scRNA-seq processing pipeline such as CellRanger or Drop-seq tools and the mapped reads in BAM format and uses them to decompose gene expression into isoform expression data (Figure 16a). SCALPEL workflow is divided into three main modules (Figure 16b). In the first module, raw sequencing data and annotation files are processed to perform bulk quantification of the annotated isoforms. These isoforms are then truncated and collapsed, giving rise to a set of distinct isoforms with different 3' ends for quantification at single-cell resolution. In the second module, scRNA-seq reads are mapped on the set of selected isoforms and reads coming from pre-mRNAs or resulting from internal priming (IP) events are discarded. In the last module, isoforms are quantified in individual cells and an isoform digital gene expression matrix (iDGE) is generated (Supplementary figure 1). The iDGE can be processed to perform downstream single-cell level analyses such as

ⁱⁱ Ake* is first author

approach helps in the assignment of UMIs to individual isoforms by considering the global transcript structure and jointly modeling the distance of the reads with the same UMI to the 3' end of the transcripts (Supplementary figure 1).

4.2.2. Benchmarking of SCALPEL

4.2.2.1. SCALPEL accurately quantifies isoforms in simulated single-cell datasets

To assess the performance of SCALPEL to quantify isoforms in scRNA-seq data, we generated synthetic single-cell isoform expression datasets. First, we simulated single-cell gene expression values for 6,000 cells belonging to two different cell populations using splatter²⁰⁷ (Figure 17a). These cells expressed a total of 6,560 genes and 12,320 isoforms including genes with changes in expression and/or isoform usage across cell populations (Figure 17b, Supplementary table 4). Using this strategy, we generated 3 datasets with different drop-out rates: one similar to a real 10x dataset used as reference and two other datasets with lower coverage (Figure 17c). Given that SCALPEL uses as input mapped reads, we used an available method, scr⁴eam (<https://github.com/plasslab/scr4eam>), to generate isoform-aware realistic scRNA-seq reads for the synthetic iDGEs. We used these three datasets to compare SCALPEL isoform quantifications with simulated isoform quantification. In all datasets, we find a high correlation between simulated isoform abundances and the isoform quantification provided by SCALPEL (Pearson correlation coefficient $r \geq 0.8$, Figure 17d-f).

4.2.2.2. SCALPEL outperforms existing tools in sensitivity and specificity for DIU identification

We used the generated synthetic datasets to benchmark the performance of SCALPEL against existing tools developed to quantify APA in scRNA-seq data^{38,79,80,150–152}. Considering the underlying quantification strategy, these methods can be divided into peak-calling based tools (Sierra, scAPA, scAPAtap, SCAPTURE, and scDaPars), and isoform quantification tools (scUTRquant) (Supplementary table 5). Following the quantification of peaks or isoforms according to the default parameters of each tool, we performed DIU analysis. In the case of scUTRquant⁷⁹, which uses an extended curated 3' UTR annotation (3' UTRome), we performed the benchmarking using both the 3' UTRome (scUTRquant) and the same annotation as the other tools (scUTRquant*).

Overall, our analyses shows a clear difference in sensitivity between peak and isoform-based methods. Peak-based methods quantified fewer genes and isoforms than isoform-based methods, which showed similar sensitivity (Figure 18a, b). Some of the sensitivity differences between peak and isoform tools can be explained by the constraints of the prediction methods. In particular, the low number of isoforms quantified by scDaPars⁸⁰ could be explained because this tool only predicts PAS in annotated 3' UTRs. As expected, sequencing depth impacts the quantification of genes and isoforms. Most methods detect less genes and isoforms in the datasets with medium and low UMI per isoform.

We next identified DIU genes across the two cell populations simulated. In all three simulated datasets, SCALPEL recovered the highest number of DIU genes

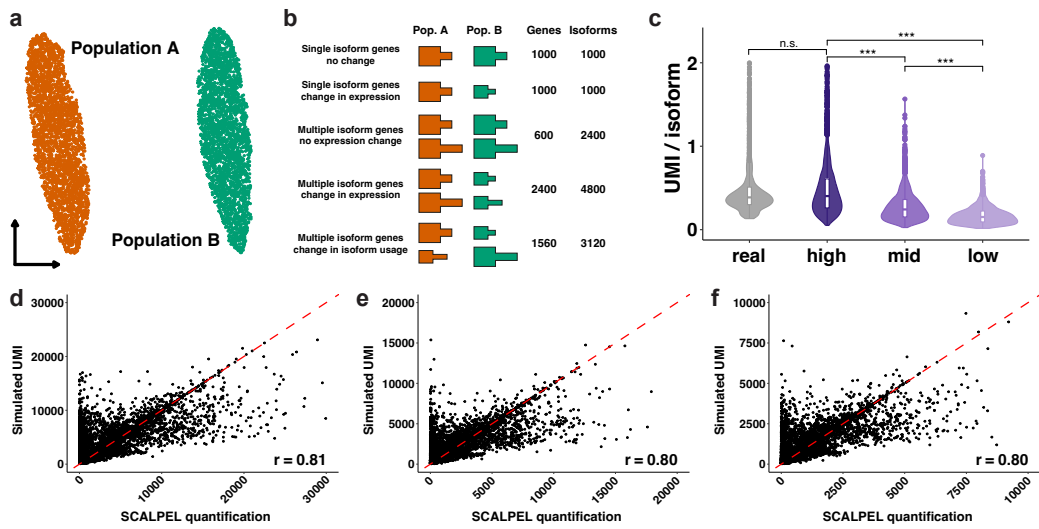


Figure 17. Benchmark of SCALPEL using synthetic data. **a.** UMAP plots depicting the two cell populations, population A (orange) and population B (green), simulated with minnow. **b.** Graphical representation of the types of genes included in the simulated data. **c.** Violin plots showing the relative number of UMIs per isoform in a reference 10x dataset (grey) and those in the simulated datasets with different dropout rates (high, mid and low depicted in dark, medium and light purple). The real dataset contains information for 40750 isoforms across 2042 cells. 12320 isoforms in 6000 cells were simulated for the high, mid and low datasets. Differences in the distribution of UMIs per isoform were tested using Mann-Whitney-Wilcoxon two-sided test. The center of the box plot is denoted by the median, a horizontal line dividing the box into two equal halves. The bounds of the box are defined by the lower quartile (25th percentile) and the upper quartile (75th percentile). The whiskers extend from the box and represent the data points that fall within 1.5 times the interquartile range (IQR) from the lower and upper quartiles. Any data point outside this range is considered an outlier and plotted individually. P-values for pairwise comparisons (Bonferroni-adjusted) are: High vs real p-value=0.86; high vs mid p-value = 4.40e-259; high vs low p-value = 2.22e-308; mid vs low p-value = 2.6e-174. **d-f.** Correlation between simulated isoform abundances (y-axis) and predicted isoform abundances (x-axis) for the high (**d**), medium (**e**) and low (**f**) sequencing depth simulated datasets.

closely followed by scUTRquant* and scUTRquant (Figure 18c, Supplementary table 8). Both methods have a superior performance than peak-based tools

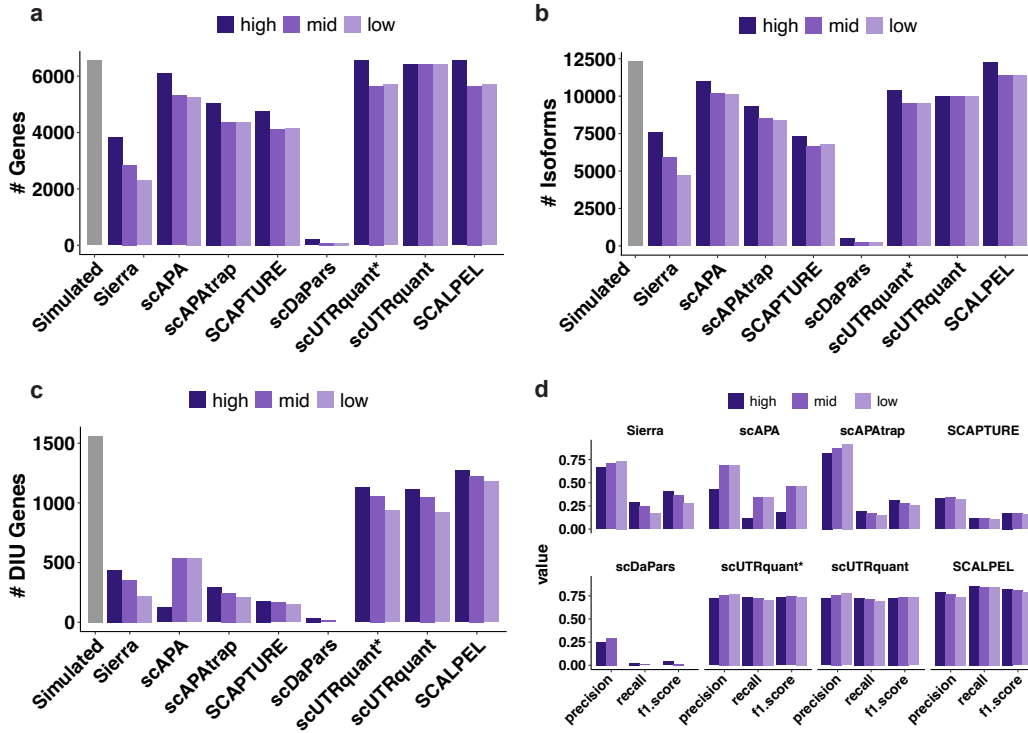


Figure 18. Number of correctly identified features. Number of genes (a), isoforms (b) and DIU genes (c) by each of the sequencing tools in the high (dark purple), medium (medium purple) and low (light purple) simulated datasets. As reference, we provide the number of simulated genes, isoforms and DIU genes simulated (grey bar). d. Summary of the performance of the different tools benchmarked on the three synthetic datasets.

although SCALPEL has higher sensitivity than scUTRquant (Figure 18d, Supplementary figure 10a-c). The higher sensitivity of SCALPEL is partly explained by a robust performance across expression ranges while all other tools misidentify lowly expressed (bottom 50% expression) DIU genes (Supplementary figure 10d-f). In the low expression dataset, SCALPEL correctly identifies 57% of DIU genes among Q1 genes, while scUTRquant and scUTRquant* identified 19% and 22% respectively (Supplementary table 8). We also noticed that true DIU genes predicted by SCALPEL have a high degree of agreement and are co-detected by at least one of the other benchmarked tools (high: 95%, mid: 93%,

low: 91%) (Supplementary figure 10g-i), indicating that tool agreement can be used as a metric to assess its performances on real datasets.

Finally, we decided to compare the performance of SCALPEL in terms of execution requirements and runtime. SCALPEL run time and memory usage is comparable or better than most of the tools analyzed, and only scUTRquant is faster and more memory efficient than SCALPEL. Yet, this can be directly linked to the presence of an already processed annotation as both memory and execution time increase when no 3' UTRome is provided to scUTRquant (scUTRquant*), (Supplementary figure 10j-l, Supplementary table 9).

4.2.2.3. DIU predictions are consistent within real single-cell datasets across benchmarked tools

4.2.2.3.1. Application on published mouse scRNA-seq data

We additionally benchmarked SCALPEL performance against the other tools using a publicly available single-cell dataset of mouse sperm cell differentiation generated using 10x Genomics platform (Supplementary figure 11a). In this dataset, SCALPEL identified 51,767 isoforms in 17,525 genes that were used for downstream analyses such as dimensionality reduction and clustering. We found 3 main cell populations using markers from a previous study¹⁹²: spermatocytes (SC), round spermatids (RS) and elongated spermatids (ES) (Supplementary figure 11b). Pairwise comparison between these populations shows again that the number of predicted DIU genes is generally higher for isoform-based approaches than for peak based methods (Supplementary figure 11c-e, Supplementary table 10). The higher number of identified DIU genes by SCALPEL is partly explained by a higher prediction among less expressed genes

(bottom 50% expression). Here, it is important to note that the number of DIU genes detected by scUTRquant using the standard gene annotation (scUTRquant*) is clearly reduced, indicating that the higher sensitivity of scUTRquant can be directly attributed to the use of an extended annotation (3' UTRome) and not to the algorithm *per se*.

Finally, we investigated the agreement in the prediction of genes with differential peak or isoform usage across tools. We observed a substantial overlap in the predictions of SCALPEL with other tools, with more than 70% of SCALPEL predictions supported by one or more tools (Supplementary figure 11f-h). SCALPEL and scUTRquant showed the highest agreement on the identified DIU genes across the cell types (Supplementary figure 11i-n).

4.2.2.3.2. Application on scRNA-seq data from hiPSCs derived cells

Furthermore, we benchmarked the performance of SCALPEL on a shallower single-cell dataset generated in house with a Drop-seq platform containing human induced pluripotent stem cells (iPSCs) and neural progenitor cells (NPCs) (Figure 19a, b). In this dataset, SCALPEL quantified 68,813 isoforms in 16,544 genes and more than 10,000 genes with two or more isoforms (Supplementary Table 6-8). When performing the benchmark on the neuronal dataset, SCALPEL and scAPAtap showed higher sensitivity in the identification of DIU genes while keeping a high degree of agreement with other tools (Supplementary figure 12a, b, c). In this case, although scUTRquant quantified a high number of isoforms (40,002) and genes with multiple isoforms (10,481), only 246 DIU genes were identified between NPCs and iPSCs (Supplementary figure 12d). This drop in the

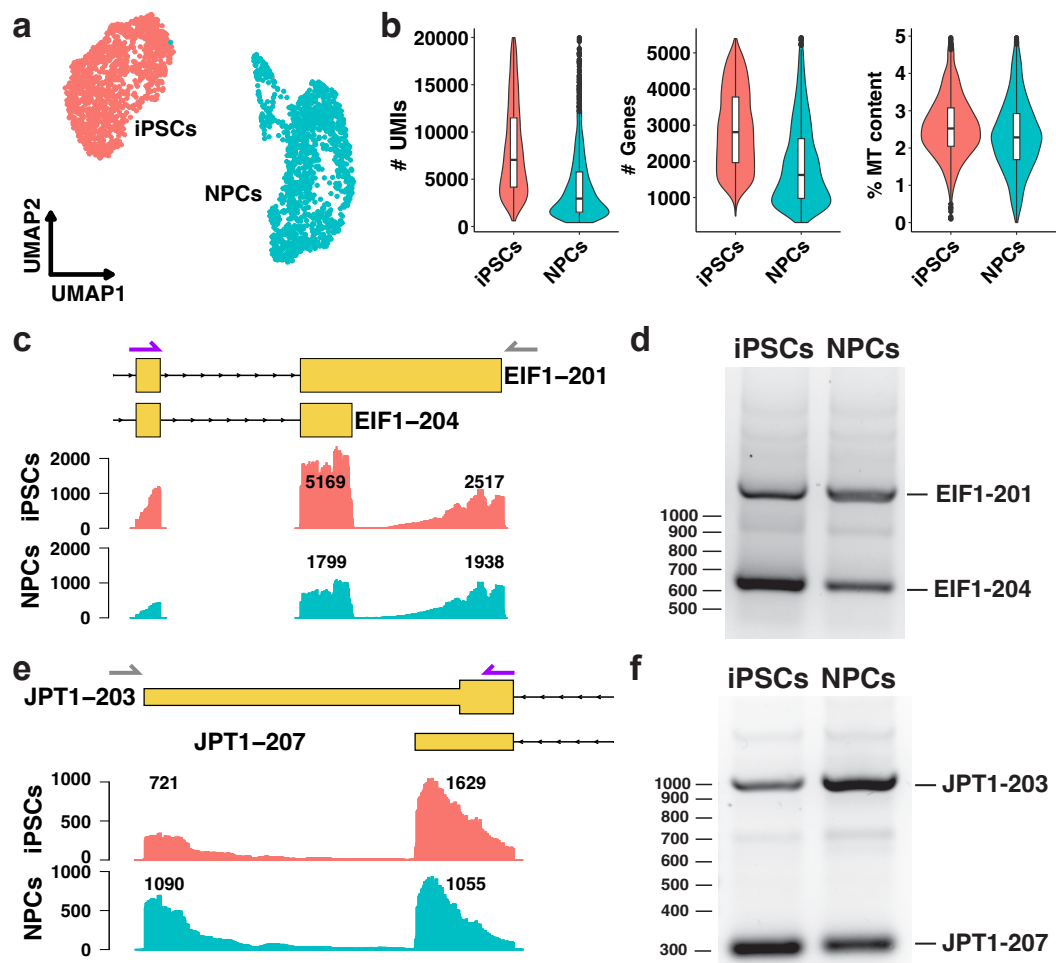


Figure 19. SCALPEL predictions on iPSC and NPC data can be experimentally validated. **a.** Clustering analysis identifies 2 main populations corresponding to iPSCs (salmon) and NPCs (cyan). **b.** Distribution of the number of UMIs, Genes and % of mitochondrial UMIs (MT) of the two samples analyze. **c, e.** SCALPEL identifies a significant change in isoform usage in EIF1 (**c**) and JPT1 (**e**) genes between iPSCs and NPCs. Coverage plots show the distribution of filtered reads along isoforms in both conditions. Relative expression of isoform abundance by SCALPEL is provided on the custom tracks. The location of the oligo dT primer (grey) and the gene-specific primer (purple) used for experimental validations are shown on top of the gene diagrams. **d, f.** Experimental validation in bulk of the isoform changes predicted by SCALPEL using 3' RACE. **d.** EIF1-204 isoform is more highly expressed in iPSCs than in NPCs while the expression of the EIF1-201 isoform is similar in both conditions. **f.** Relative abundance of JPT1-203 in NPCs is higher than in iPSCs while JPT1-207 is more similar across condition.

number of detected DIU genes is likely arising from stringent default parameters which discard genes expressed in a few cells (*minCellsPerGene*=50).

Overall, the results from the benchmark show that SCALPEL has higher sensitivity than all other tools while keeping a good precision and performance in real datasets generated with 10x and Drop-seq technology.

4.2.2.4. Experimental validation confirms DIU changes predicted by SCALPEL

Considering that iPSCs and NPCs are easily distinguished experimentally, we decided to use this dataset to experimentally validate SCALPEL predictions. To this end, we selected five genes predicted to have changes in isoform usage by SCALPEL across cell types (Supplementary table 11) and validated the expression of the isoforms quantified by SCALPEL in NPCs and iPSCs using 3' RACE. With this analysis, we confirmed the 3' UTR lengthening of two DIU genes predicted by SCALPEL, *EIF1* and *JPT1*. In both examples, SCALPEL quantified two isoforms differing in the length of the last exon and predicted a shift in the expression of the short and the long isoform between iPSCs and NPCs (Figure 19c, d) that was also validated with the 3' RACE (Figure 19e, f). For the other 3 genes tested, we were able to detect all isoforms predicted by SCALPEL in two of the remaining genes (Supplementary figure 13).

4.2.3. Isoform-based clustering enables the identify a novel cell population

We decided to use SCALPEL isoform quantification to investigate the benefits of using isoform-based analyses instead of gene-based analyses for clustering and

downstream analyses. Using the mouse spermatogenesis dataset, we observed that the use of isoforms instead of genes for clustering analysis results in the identification of the same three main cell populations (Figure 20a; 94% agreement in cell assignment).

To investigate if isoforms provide additional biological insights in single-cell analyses, we decided to perform a high-resolution clustering analysis of the spermatogenesis dataset. As expected, at higher resolutions both gene and isoform-based analyses identify more cell populations while keeping a high degree of similarity across clustering solutions (Figure 20b, Supplementary figure 3a, Supplementary table 12, Supplementary table 13). At this higher resolution, we identified a new cell population of RS cells, RS6, in the isoform-based analysis that could not be distinguished using gene-based clustering. This population is composed of cells from other RS cell clusters and not the result of subclustering existing RS cell populations (Figure 20b, Supplementary figure 3a, b). Compared to other RS populations, the RS6 shows few changes in gene expression (Supplementary figure 3c). Yet, RS6 cluster presents a clear set of isoform markers that cannot be identified in the gene-based analysis (Supplementary figure 3d, Supplementary table 13). GO term enrichment analysis using RS6 isoform markers identified biological processes associated to cilium organization and organelle assembly (Figure 20c, Supplementary table 13), which are essential processes for the differentiation and maturation of RS cells^{208,209}. DIU analysis across RS populations identified 543 genes with differential isoform usage in RS6 cells (Supplementary table 15). These isoforms are not the result of changes in gene expression as only 3 DIU genes are differentially expressed.

Among them we identified *Dnah3*, a structural component of the axoneme essential for the flagellum assembly and the acquisition of sperm motility, and *Msi2*, an RNA binding protein that play a central role in coordinating post-transcriptional programs essential for late spermatocytes to early spermatid differentiation^{210–212}. These genes exhibited higher expression of longer transcript isoforms in RS6 compared to neighboring RS cells, suggesting that these changes are affecting the resulting proteins and their function within the different cell types (Supplementary figure 3e, f).

Together, these results suggest that RS6 cell population corresponds to elongating spermatids²¹³, an intermediate population state between RS and ES populations previously described morphologically in the literature that had not been identified using single-cell gene expression profiles.

4.2.4. Investigation of APA profiles and 3' end isoform variability across the cell types

4.2.4.1. Isoform quantification recapitulates a global 3' UTR shortening during mouse spermatogenesis

In this dataset, most of the genes express a single transcript in individual cells. Yet, we detect on average 917 genes with two or more isoforms per cell, which may be potentially regulated across cell types (Supplementary figure 14a). Given that isoform quantification may be useful to identify new cell populations, we investigated how informative is isoform expression to define cell type identity. For this purpose, we computed for each gene its information content defined as the sum of the information from all its expressed isoforms²¹⁴. The higher the information content of a gene, the more cell-type-specific the expression of its isoforms is. Using this approach, we noted that most genes had clear cell type specific expression bias (Figure 21a). Manual inspection showed that in many cases all isoforms from the same gene showed similar expression changes across cell types, indicating that gene information content mainly reflects transcriptional changes (Figure 21b). Thus, we used a chi-square test to identify genes in which isoform usage changes across conditions, reflecting a regulation at the post-transcriptional level. Using this approach, we identified 4,196 genes displaying changes in isoform usage across cell types (Figure 21a, red dots; Supplementary table 16). Pairwise DIU analysis showed that most of these genes present changes in isoform usage independent of changes in gene expression, indicating that those events are only regulated at the post-transcriptional level (Figure 21c).

In most cases, changes in APA across cell types involved changes in tandem APA sites, which only affect the length of the 3' UTR, or complex events including isoforms with changes in the 3' UTR length and the exon composition (Figure 21d). These results indicate that the main effect of APA during spermatogenesis is the regulation of the 3' UTR length. One of these genes is *Smg7*, a gene that plays a crucial role in male germ cell differentiation in mice through its role in nonsense-mediated mRNA decay²⁵. We observed a switch in isoform usage during cell differentiation where the long isoform of *Smg7* gene (*Smg7*-202) is progressively replaced by a shorter isoform (*Smg7*-203) (Figure 21e).

Previous studies have shown that APA results in global 3' UTR shortening during sperm cell differentiation^{47,215}. Thus, we used SCALPEL predictions to assess if the observed changes in 3' UTR length reflected a coordinated shortening during sperm cell differentiation. To this end, we used the isoform quantification data to order cells by pseudotime and calculated the average 3' UTR length per cell. In agreement with previous studies, we observed that overall, 3' UTR length shortens while cells differentiate (Figure 21f), indicating that SCALPEL predictions recapitulate known changes in 3' UTR length during mouse spermatogenesis.

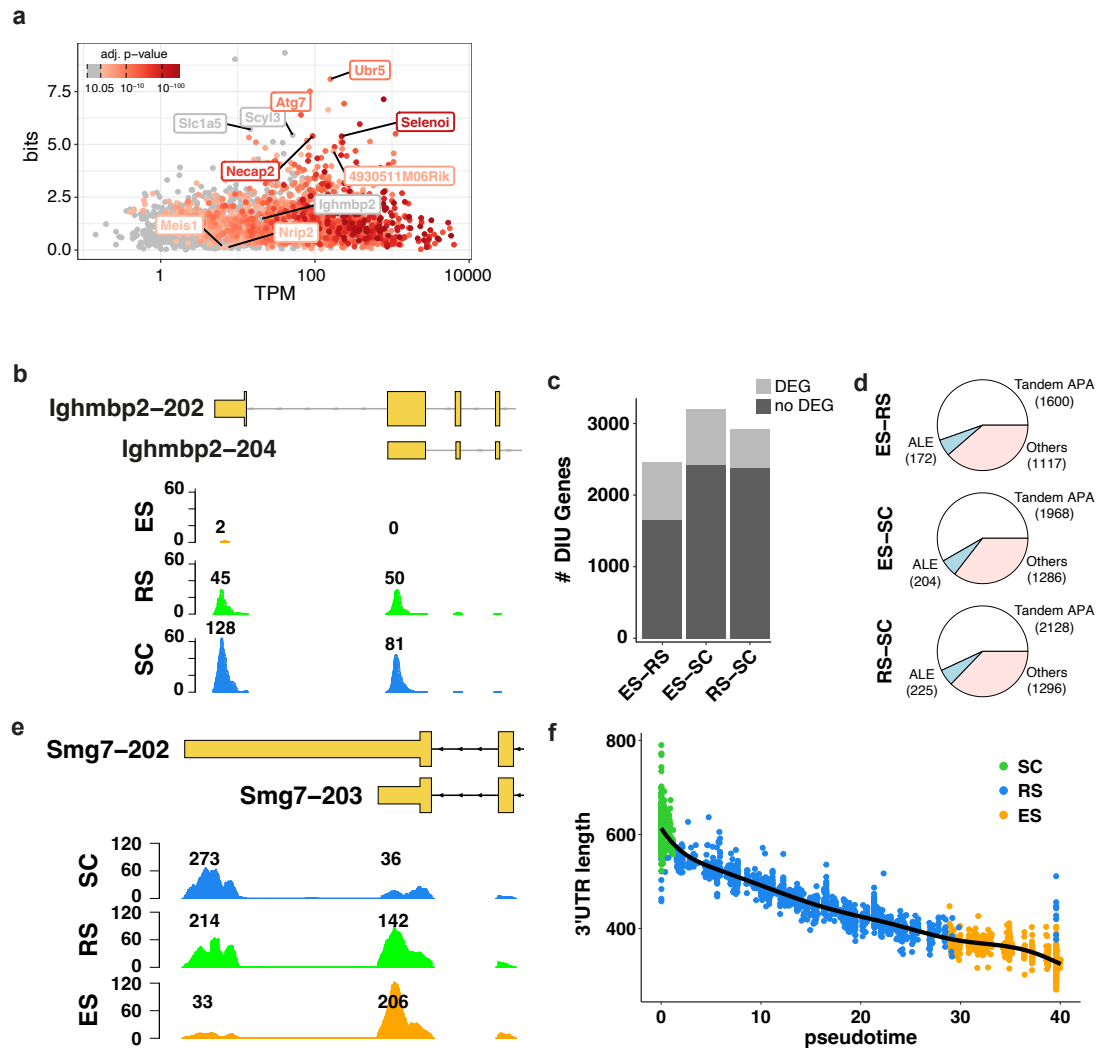


Figure 21. Isoform quantification recapitulates 3' UTR shortening during mouse spermatogenesis. **a.** Scatter plot showing the cell-type expression specificity relative to gene expression for all genes with only two expressed isoforms. The higher the information content of a gene, the more cell-type specific its expression is. Colored dots represent genes whose isoform expression usage changes across conditions (Chi-squared test adjusted p-value < 0.05). **b.** SCALPEL quantification of isoform usage from *lghmbp2*, a gene with high cell type specificity identified in **(a)** whose isoforms show the same relative usage across cell types during mouse sperm cell differentiation. Coverage plots show the distribution of filtered reads along isoforms. SCALPEL quantifications of the different isoforms are shown on top of the custom tracks. **c.** Number of DIU genes identified in each pair of samples that are also differentially expressed (light grey) or not (dark grey). Most genes show changes in isoform usage independently of changes in expression. **d.** Classification of DIU

genes depending on the type of APA event identified. **e.** Quantification of *Smg7* isoforms by SCALPEL. *Smg7* gene has low cell type specificity but shows significant changes in isoform usage across conditions. SCALPEL quantification and coverage plots show a gradual switch in isoform usage during the differentiation from SC to ES cells. **f.** Pseudotemporal ordering of cells confirms the overall shortening of 3' UTRs during mouse sperm cell differentiation.

4.2.4.2. Isoform biotypes remain stable during APA changes

Finally, we investigated if the changes in isoform usage across cell types involved changes in the types of transcripts (i.e., the biotype) expressed in the cells that could further alter cell function by promoting, for instance, the degradation of expressed mRNAs or the production of non-coding isoforms of a gene. We observed that in ~80% of the cases, changes in APA did not involve a change in the type of transcripts expressed (Supplementary figure 14b). In the cases where the biotype of the most expressed isoform changed across cell types, in 82% of the cases this change involved a switch from a mRNA to a non-coding or NMD isoform or *vice versa* (ES-RS: 429, RS-SC: 406, ES-SC: 690; Fig. S7c). In these cases, APA could alter significantly protein levels and lead to stronger functional changes.

4.2.4.3. Isoform biotypes remain stable during APA changes

Finally, we investigated if the changes in isoform usage across cell types involved changes in the types of transcripts (i.e., the biotype) expressed in the cells that could further alter cell function by promoting, for instance, the degradation of expressed mRNAs or the production of non-coding isoforms of a gene. We observed that in ~80% of the cases, changes in APA did not involve a change in

the type of transcripts expressed (Supplementary figure 14b). In the cases where the biotype of the most expressed isoform changed across cell types, in 82% of the cases this change involved a switch from a mRNA to a non-coding or NMD isoform or *vice versa* (ES-RS: 429, RS-SC: 406, ES-SC: 690; Fig. S7c). In these cases, APA could alter significantly protein levels and lead to stronger functional changes.

4.2.4.4. Isoform expression changes reflect miRNA function

Since it is known that APA changes significantly during neurogenesis²¹⁶, and that miRNA regulation is very important in this process^{191,217,218}, we investigated if changes in isoform usage in NPCs compared to iPSCs were driven by miRNA function. To address this question, we downloaded the predicted miRNA target sites on the human genome from the MBS database¹⁸⁹ and identified all isoforms

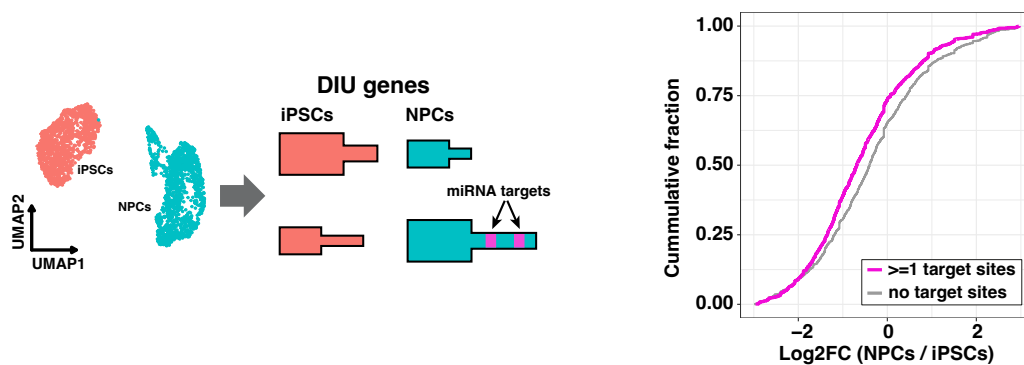


Figure 22. SCALPEL predictions on iPSC and NPC reflect miRNA function.

Cumulative distribution plot of log2fcs showing the difference in the expression of isoforms from DIU genes where one of them contains miR-128-5p target sites (pink) and the other does not (grey), indicating that changes in isoform usage between iPSCs and NPCs can be attributed to miRNAs known to be implicated in neurogenesis.

targeted by miRNAs previously associated to NPCs¹⁹¹ (Supplementary table 10). Given that isoforms with different 3' ends could contain different regulatory elements such as miRNA target sites¹, and that miRNAs usually downregulate their target RNAs²¹⁹, we compared the fold change distribution of isoforms containing target sites of miRNAs expressed in NPCs with those of non-targeted isoforms from the same genes (Figure 22). This analysis identified significant downregulation of isoforms targeted by let-7b, miR-9, miR-124, miR-128, miR-153, miR-199a, and miR-34 in NPCs compared to iPSCs (FDR < 0.05; Figure 22 and Supplementary table 17). This result suggests that miRNAs can explain some of the isoform expression changes predicted by SCALPEL during the differentiation of iPSCs to NPCs.

4.2.5. SCALPEL improves isoform quantification of novel isoforms predicted using single-cell long read sequencing

We investigated if SCALPEL could be used to improve the quantification of isoforms predicted in individual cells using long-read single-cell sequencing data. We reanalyzed two single-cell datasets from P28 mouse hippocampus and visual cortex containing 10x 3'seq scRNA-seq data and paired PacBio data⁷¹ (Figure 23a, Supplementary figure 15a, Supplementary table 18). We performed standard single-cell analysis using gene expression data and subset the data of the main five cell populations: excitatory neurons, astrocytes, NPCs, oligodendrocytes and microglia. As expected, comparison of cell assignments using gene or isoform-based quantification revealed a high degree of agreement between the two approaches (Figure 23a, Supplementary table 19, 20). Given that SCALPEL can only distinguish isoforms with changes at the 3' end, we

collapsed isoforms and kept only those with changes in the last 600 nt. Then, we compared the abundance per cluster of these isoforms according to isoquant and SCALPEL. The quantification by SCALPEL clearly correlates with that of isoquant although the expression range of SCALPEL is broader as the quantification is done using the single-cell short read data and not the long-read data (Figure 23b, Supplementary figure 15b). We also ran SCALPEL to identify genes with changes in isoform usage across cell types. In this dataset, we found that most changes in isoform usage happen between glial cells (microglia, astrocytes, and oligodendrocytes) and the excitatory neurons (Supplementary table 21). Among these genes, we identified examples in which isoform switch events predicted by SCALPEL were supported by the long-read data. One of those examples is the gene *Cdc42*. SCALPEL detects a change in isoform usage between excitatory neurons and microglia/astrocytes, which is validated with the PacBio data (Figure 23c). *Dclk1* is another example in which SCALPEL predicts a change in isoform usage between astrocytes/microglia and excitatory neurons. In this case, microglia and astrocytes express only a short version of the gene while excitatory neurons express a much longer isoform (Figure 23d). Here, it is interesting to highlight that isoquant quantifications disregard changes in the 3' ends and focus on changes in exon junctions. While SCALPEL quantifies two isoforms with tandem poly(A) sites (*Dclk1-202* and *Dclk1-208*), isoquant focuses on the changes in exon composition. Together, these analyses demonstrate that SCALPEL can be used to quantify isoform diversity derived from paired long and short read data with single-cell resolution and that provides additional information to the long-read data that disregards changes in the 3' ends that do not affect exon composition.

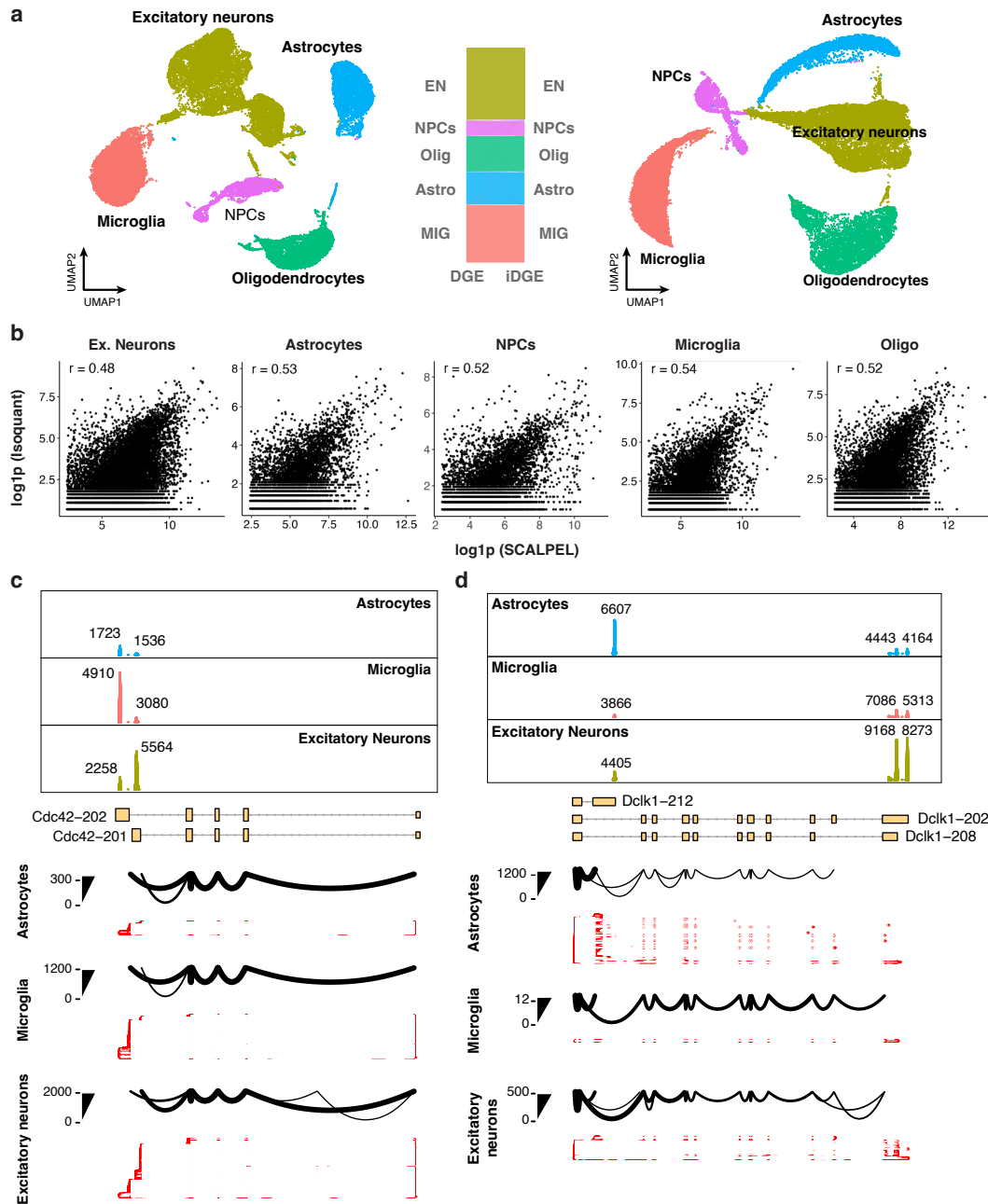


Figure 23. SCALPEL provides isoform quantification on individual cells using paired short and long scRNA-seq datasets. **a.** UMAP plots depicting the main 5 cell populations selected from Jogleklar et al. identified using gene and isoform-based quantifications. **b.** Comparison of isoform expression quantification for each cell type using isoquant (PacBio data) and SCALPEL. **c, d.** SCALPEL identifies changes in isoform usage in *Cdc42* (**c**) and *Dclk1* (**d**) between astrocytes/microglia and excitatory neurons. Custom tracks on top show the mapping of filtered short reads and SCALPEL quantification for each cell type. On the bottom, sashimi plots show the relative usage of different isoforms in individual cell types and the aligned long-read data supporting them.

5. Discussion

Single-cell transcriptomics methods have become a powerful and increasingly standard approach for studying transcriptomic changes across diverse samples and conditions at single-cell resolution. These technologies are relatively recent compared to bulk transcriptomics methods such as RNA-seq or 3' seq. As a result, standardized protocols for sample preparation are still lacking in many contexts. In this thesis, we aimed to characterize alternative polyadenylation (APA) at single-cell resolution using 3'-tagged scRNA-seq data.

To achieve this, we structured the work around two main objectives:

First, we optimized single-cell transcriptomics protocols for neural and glial cell populations derived from human induced pluripotent stem cells (hiPSCs) by systematically evaluating how commonly used preservation methods, including DMSO cryopreservation, methanol fixation, and commercial reagents such as ACME and vivoPHIX affect transcriptomic data quality, cell composition, and gene expression profiles. These evaluations allowed us to ensure the reliability and interpretability of scRNA-seq data for APA analysis, particularly given its sensitivity to the technical noise.

Second, we developed and benchmarked SCALPEL, a computational pipeline designed to quantify transcript isoforms from 3'-tagged scRNA-seq datasets, enabling the detection and analysis of APA events at single-cell resolution.

5.1. Objective 1: Optimize single-cell transcriptomics protocols to work with iPSC-derived neuron and glial cells

In the first objective of our work, we investigated how sample preservation methods employed during the library preparation for single-cell experiment impact the library complexity, changes in cell composition and alterations in the expression profile of individual cells.

5.1.1. Sample preservation methods impact the library quality in single-cell experiments

Fixation method	Library yield	Library complexity	Low-quality cells	Composition bias	Stress signature	Apoptosis signature	Expression bias
DMSO	HIGH	HIGH	FEW	STRONG	HIGH	MEDIUM	LOW
Methanol	HIGH	HIGH	FEW	WEAK	MEDIUM	LOW	LOW
ACME	LOW	LOW	MANY	WEAK	MEDIUM	MEDIUM	MEDIUM
vivoPHIX	LOW	LOW	MANY	WEAK	MEDIUM	MEDIUM	MEDIUM

Table 2: Summary of the effects of preservation methods on single-cell libraries

Our results demonstrate that preservation/fixation methods significantly impact the quality of single-cell RNA-seq libraries (Table 2). Key indicators of library complexity, such as cDNA yield, the number of genes and UMIs detected per cell, and the proportion of low quality cells varied substantially across preservation methods. Methanol-fixed and DMSO-cryopreserved samples showed a complexity profile comparable to those of Fresh samples, with a high cDNA yield and median gene/UMI counts per cell. In contrast, samples fixed with ACME and vivoPHIX-fixed showed a marked reduction in data quality.

These impairments were also associated with a marked decrease in the number of detected UMIs and genes per cell under the same conditions. In the case of vivoPHIX, lower initial RNA integrity values (RIN) may partially explain the poor results. In contrast, ACME samples, despite having initial RIN values comparable to those of Fresh samples prior to fixation, likely underwent degradation during or after the fixation process. This observation aligns with previous reports showing that RNA integrity declines over time in ACME-fixed samples resulting in reduced library complexity¹³⁶. These effects translated into higher dropout rates, leading to a larger proportion of low-quality cells being filtered out during data preprocessing, and reaching up to 49% in vivoPHIX samples, compared to fewer than 10% in Fresh, DMSO, and methanol-fixed conditions. The reduced library complexity of these samples due to a higher dropout rate may have introduced biases in cell population clustering (Figure 14) and contributed to the number of differentially expressed genes (DEGs) identified in each cluster (Figure 15c).

5.1.2. DMSO cryopreservation alters cell-type composition

Beyond differences in library quality, our results showed that preservation methods can introduce substantial biases in cell population composition. Notably, DMSO cryopreserved samples exhibited a pronounced depletion of excitatory neuron populations compared to Fresh and methanol-fixed samples (Figure 13). Moreover, matched comparisons on DMSO and Methanol samples derived from the same differentiation experiment (M3, M4, D1, D2) confirmed that the neuron depletion observed in DMSO treated samples was attributable to the preservation method rather than biological variability (Supplementary figure 6).

Interestingly, these compositional changes were not observed in samples preserved with ACME, or vivoPHIX, suggesting that cryopreservation may selectively impair fragile neuronal populations. While DMSO has been previously reported as an effective method of cell preservation for single-cell transcriptomics^{101,142}, recent studies highlighted the poor tolerance of mature neurons to the effect of DMSO^{220,221}. A study by Li et al. evaluating the impact of cryopreservation across different stages of neuron differentiation from human hiPSCs showed that mature neurons were the most sensitive to DMSO exposure²²². This reduction in the numbers of neurons could be attributed to DMSO pre-freeze toxicity, freeze-induced injury and functional impairment post-thaw. The latter could explain the higher expression of stress genes in DMSO samples compared to Fresh and all other fixed samples. Yet, the obtained neurons do not show strong expression biases as they often cluster with Fresh and Methanol-fixed samples and show very few significant DEGs (Figure 14). While our results indicate that DMSO is not suitable for compositional analysis of hiPSC-derived cells, it could be used for profiling pure neural samples when cell availability is not an issue, or other cell types that are less sensitive to cryopreservation-induced stress.

5.1.3. Preservation methods induce stress and apoptotic signatures

In addition to affecting library complexity and cell-type composition, our analyses revealed that different fixation/preservation methods can alter the overall gene expression profile of cell clusters. This was particularly evident in vivoPHIX and ACME fixed-samples, which showed global expression shifts (Figure 14,

Supplementary figure 8, 9) as well as cell-type specific changes in gene expression (Figure 15c, Supplementary table 3). Quantification of immediate early genes (IEGs) and apoptotic markers revealed upregulation in all vivoPHIX, ACME, and DMSO treated samples compared to Fresh, with the strongest induction observed in DMSO samples. These results are consistent with previous reports linking higher IEG expression to dissociation and cryopreservation biases^{135,139}, and highlight the importance of accurate controls and validation experiments to confirm gene expression changes observed in single-cell data.

Additionally, we investigated whether the fixation methods could affect the detection of DEGs across the cell populations, given their impact on library complexity and the number of genes and UMIs detected by cells. Since DEG detection is directly related to the sequencing depth and transcript capture efficiency, subtle changes in gene expression due to biological or experimental conditions may be missed in low complexity libraries. DEG analysis comparing each preservation sample to the Fresh samples revealed a large number of DEGs in vivoPHIX samples and a moderate number in ACME-fixed cells (Figure 15c). These effects are likely driven by the lower gene and UMI counts per cell observed in these conditions (Figure 11c), and increasing the number of sequenced cells could compensate for this limitation. In contrast, DMSO and Methanol conditions yielded relatively few DEGs across cell-types. Moreover, methanol-fixed cells have been successfully used in previous studies to detect changes in gene expression in single-cell data across diverse biological conditions^{223,224}. These findings indicate that Methanol fixation is the most suitable approach, not only to identify homeostatic gene expression but also to identify more subtle biological differences.

Altogether in the first objective of our work, we demonstrated that preservation protocols not only alter the library quality and cell-type composition, but also affect gene expression profiles in a cell-type-specific manner. Our comparative analysis, based on multiple quality metrics (Table 2) indicates that methanol fixation provides the best preservation method for single-cell transcriptomics of hiPSC-derived neural cells. Methanol-fixed libraries closely resemble Fresh samples in terms of library complexity and do not introduce strong biases in gene expression that affect the overall transcriptomic profile of cells or cell composition, making them the most suitable approach for transcriptomic analysis.

Although we evaluated each preservation condition using a limited number of replicates (2-4), this sample size aligns with current experimental standards in single-cell studies, and is considered acceptable regarding the cost of individual single-cell experiments. Nevertheless, this limitation restricts our ability to fully assess the influence of all potential confounding variables. Factors such as preservation duration, cell line variability, differentiation protocol, and reagent batch effects (e.g., bead lots) can all impact the final transcriptomic output. While these variables were not systematically explored in this study, they represent important considerations for future work. Our findings underscore the necessity of optimizing sample processing protocols for each experimental context, as even subtle differences in handling can significantly affect the outcomes of single-cell transcriptomics experiments.

5.2. Objective 2: Implementation of a computational method for characterization of APA with scRNA-seq datasets

5.2.1. SCALPEL, a computational framework for isoform quantification using single-cell data

In this second objective of our work, we aimed to quantify isoform diversity at single-cell resolution using conventional 3' tagged scRNA-seq data. For this purpose, we developed SCALPEL, a new tool for sensitive isoform quantification and visualization using conventional 3' based scRNA-seq data. Implementation of SCALPEL within Nextflow provides a robust framework that enables reproducibility, portability, and efficient resource management across diverse computing environments²⁰⁶. This strategie for building scalable bioinformatic pipelines is increasingly adopted within the bioinformatic community, driven by the growing need for reproducible and portable analyses across diverse research lab and computational setups^{225,226}.

5.2.2. SCALPEL accurately quantifies isoform and outperform published tools for APA characterization in single-cell data

Comparison of SCALPEL to existing other APA quantification tools, including both peak calling and isoform-centric methods using synthetic data, showed that it correctly identifies a higher fraction of simulated genes and isoforms, and has a high accuracy in the detection of DIU genes (Figure 18a, b, c, d). Its performance remained stable even under high dropout conditions, demonstrating robustness across different sequencing depths and noise levels. Importantly, benchmarking of SCALPEL on real data show that SCALPEL DIU predictions have a high

agreement with that of the other benchmarked tools (Supplementary figure 10). This approach of using the consistency of prediction across the tools as a metric to assess prediction reliability has been used in prior benchmark studies as a proxy for accuracy ¹⁴⁹.

Additionally, we also show that SCALPEL performance remains similar in the 10x dataset as in the Drop-seq dataset where other tools tend to underperform, highlighting its broad applicability across platforms (Supplementary figure 11, 12). Moreover, we showed that SCALPEL predictions can be validated experimentally using 3' RACE (Figure 19d, f, Supplementary figure 13) and long read sequencing (Figure 22c, d) data. While 3' RACE is not quantitative and can only be used to validate specific isoforms in bulk, we consider that the validation of some of SCALPEL predictions using this method confirms the reliability of the predictions. Similarly, while PacBio only provides reliable cell-type specific measurements, the significant correlation between isoquant and SCALPEL using pseudobulk data supports SCALPEL predictions.

5.2.3. SCALPEL isoform quantification conserves cell-type specificity and enable to recover novel cell types

Importantly, we demonstrated in our results, that the iDGE provided by SCALPEL can be used to perform standard single-cell analysis such as dimensionality reduction, clustering and pseudotime analysis. At low resolution, SCALPEL iDGE provides very similar clustering solutions than gene-based analyses (Figure 20a). Furthermore, at higher resolutions we showed that isoform quantification reveals additional granularity in transcriptomic landscapes, allowing the identification of

novel cell states that remain undetected by gene-level data alone (Figure 20b, Supplementary figure 3b), indicating that we can capture post-transcriptional regulatory events by quantifying isoforms in individual cells. Recent isoform based studies enabled to uncover novel cell states that cannot be fully uncovered with gene-level data^{227,228}. For instance, Sommerkamp et al. profiled hematopoietic stem cells (HSCs) during stem cell activation and revealed transcriptome-wide dynamic APA patterns upon homeostatic or stress-induced transition from quiescence to proliferation. Particularly, they showed that quiescent HSCs harbor longer 3' UTRs and extensive intronic PAS usage, whereas activated HSCs and progenitors exhibited widespread 3' UTR shortening, notably in metabolic regulators²²⁹. Similarly, Patowary et al. leveraged single molecule long-read sequencing to deeply profile the full-length transcriptome of developing human neocortex at single-cell resolution, and identify previously uncharacterized cell states using isoform-based single-cell clustering²²⁷. Together, these APA-derived signatures constitute an additional regulatory layer that enables the identification of cell states shaped by post-transcriptional programs, such as 3'UTR remodeling which remain obscured in conventional gene-level analysis.

5.2.4. SCALPEL identifies global APA dynamic and isoform usage change across cell-types

Using SCALPEL, we identified widespread changes in APA and isoform usage across cell types, revealing both global and cell-type-specific isoform usage. In the mouse spermatogenesis dataset, SCALPEL accurately recapitulated the previously reported trend of global 3' UTR shortening^{47,215}, mainly driven by changes in the use of tandem APA sites (Figure 20f, Supplementary figure 14b, c). In the human iPSC & NPC data, changes in isoform quantification reflect

miRNA function at the single-cell level, as changes in isoform usage across cell types can be directly linked to the presence of cell type specific miRNAs (Figure 21). By integrating miRNA target site predictions and SCALPEL isoform quantification data, we found that isoforms harboring target sites for NPC-enriched miRNAs (e.g., let-7b, miR-9, miR-124) were significantly downregulated compared to their non-targeted isoforms. This observation supports the hypothesis that the change in isoform usage with distinct 3' UTR length results from miRNA-mediated repression, a mechanism known to be critical during neurogenesis^{42,50}. Interestingly, analysis of isoform biotype across cell types showed that APA rarely results in a switch in transcript biotype. In about 80% of cases, APA did not alter the classification of transcripts. This observation is consistent with previous studies showing that most APA events occur within the 3' UTR, leaving the coding sequence mainly unchanged³.

Finally, we also show that SCALPEL can be used to improve isoform quantification obtained from paired long and short-read scRNA-seq datasets and accurately quantifies isoform changes across cell-types (Figure 22). In these cases, SCALPEL can be exploited to quantify the relative abundance of isoforms predicted using long-read data at the individual cell level, which cannot be done with standard scRNA-seq analysis pipelines. Additionally, it provides the quantification of shortening/lengthening events in the last exon, i.e. tandem APA sites, which are usually not considered using long-read data that focuses on events that change exon composition.

Altogether, in this second objective of our work, we completed the development of SCALPEL, a computational tool designed to quantify isoform expression in

single-cell RNA-seq data. This pipeline environment allowed it to be fully reproducible, portable, and easy to execute across computational environments. A limitation of SCALPEL is that it relies on genome annotations and therefore cannot perform de novo identification of PAS, unlike tools such as scAPAtap that functions in a genome unaware manner¹⁵¹. Importantly, our results, obtained using short-read-based inference of isoform expression through SCALPEL, similarly demonstrate that isoform quantification can expand the single-cell analysis resolution and reveals the transcript diversity, even in the absence of long read data. While the usage of 3' tagged single-cell RNA-seq dataset by SCALPEL, making it inherently more sensitive to detecting APA events; It is not only restricted to processing reads mapping within the 3' UTR of the isoforms but rather operates from a distance scope (600bp) from their 3' end coordinates, increasing its accuracy in deciphering broader post-transcriptional changes. Furthermore, this study focused on a limited number of datasets and expanding the application of SCALPEL to a broader range of biological systems will be necessary to fully validate its utility across different contexts.

In summary, we achieved the objective of developing a new tool for sensitive isoform quantification and APA characterization at single-cell resolution, by addressing both experimental and computational challenges inherent to 3'-tagged scRNA-seq. Through a systematic evaluation of preservation methods in hiPSC-derived neural populations, we demonstrated that methanol fixation provides the most reliable balance between transcriptome integrity and unbiased cell-type representation, thereby offering a robust protocol for single-cell studies of sensitive cell types. We further showed that SCALPEL preserves cell-type specificity, enables the identification of novel cell states, and captures dynamic APA regulation across cell populations. Benchmark analyses demonstrated that SCALPEL outperforms several existing

computational approaches in sensitivity and specificity for detecting DIU genes across both low and high-resolution single-cell datasets. In addition, we validated SCALPEL predictions experimentally through 3'RACE and computationally using long-read sequencing data. Together, this work expands the scRNA-seq toolkit to include isoform-level analyses, thereby enabling the exploration of post-transcriptional regulation across species, tissues, and technologies, and advancing our understanding of gene regulation from the bulk to the single-cell level.

6. Conclusions

- Methanol fixation and DMSO cryopreservation fixations maintain high library complexity comparable to Fresh sample. In contrast, ACME and vivoPHIX methods results in low data quality with higher degradation and increased dropout rates.
- Cell composition is altered by DMSO preservation, and specifically depletes the neuron population.
- ACME and vivoPHIX introduce expression biases and higher stress and apoptosis signatures compared to Fresh cells. Methanol fixed cells closely resemble fresh cells in gene expression profiles and show minimal biases in differential gene expression.
- SCALPEL enables accurate, reproducible and fast isoform quantification at single-cell resolution using standard 3' tagged scRNA-seq data.
- Benchmarking analysis on synthetic and real datasets demonstrated that SCALPEL outperforms currently published tools for APA study with single-cell data in term of sensitivity and precision, and accurately predict isoform usage changes in both simulated and real single-cell datasets.
- Isoform-level clustering using SCALPEL recapitulates known cell types and reveals novel subpopulations that are not captured by gene-level analysis.

- SCALPEL enabled the identification of differential isoform usage across cell types and accurately recapitulates a global 3' UTR shortening event in cells during mouse spermatogenesis.
- SCALPEL identified changes in isoform usage between the iPSCs and NPCs that were validated experimentally using 3' RACE. and isoform shifts linked to miRNA targeting during neuronal differentiation. SCALPEL identify isoform shift linked to mRNA targetting during neuronal differentiation.

7. References

1. Tian, B. & Manley, J. L. Alternative polyadenylation of mRNA precursors. *Nature Reviews Molecular Cell Biology* 2016 18:1 **18**, 18–30 (2016).
2. Yeh, H.-S. & Yong, J. Alternative Polyadenylation of mRNAs: 3'-Untranslated Region Matters in Gene Expression. *Mol Cells* **39**, 281–285 (2016).
3. Gruber, A. J. & Zavolan, M. Alternative cleavage and polyadenylation in health and disease. *Nature Reviews Genetics* 2019 20:10 **20**, 599–614 (2019).
4. Hoque, M. *et al.* Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing. *Nat Methods* **10**, 133–139 (2013).
5. Tian, B. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res* **33**, 201–212 (2005).
6. Elkon, R., Ugalde, A. P. & Agami, R. Alternative cleavage and polyadenylation: extent, regulation and function. *Nat Rev Genet* **14**, 496–506 (2013).
7. Mitschka, S. & Mayr, C. Context-specific regulation and function of mRNA alternative polyadenylation. *Nature Reviews Molecular Cell Biology* 2022 23:12 **23**, 779–796 (2022).
8. Lu, F., Gladden, A. B. & Diehl, J. A. An Alternatively Spliced Cyclin D1 Isoform, Cyclin D1b, Is a Nuclear Oncogene¹. *Cancer Res* **63**, 7056–7061 (2003).
9. Solomon, D. A. *et al.* Cyclin D1 Splice Variants. *Journal of Biological Chemistry* **278**, 30339–30347 (2003).

10. Tian, B. & Graber, J. H. Signals for pre-mRNA cleavage and polyadenylation. *Wiley Interdiscip Rev RNA* **3**, 385–396 (2012).
11. Ren, F., Zhang, N., Zhang, L., Miller, E. & Pu, J. J. Alternative Polyadenylation: a new frontier in post transcriptional regulation. *Biomark Res* **8**, 67 (2020).
12. Boldyreva, L. V. *et al.* Fine gene expression regulation by minor sequence variations downstream of the polyadenylation signal. *Mol Biol Rep* **48**, 1539–1547 (2021).
13. Chen, W. *et al.* Alternative Polyadenylation: Methods, Findings, and Impacts. *Genomics Proteomics Bioinformatics* **15**, 287–300 (2017).
14. Zhang, Y. *et al.* Alternative polyadenylation: methods, mechanism, function, and role in cancer. *Journal of Experimental and Clinical Cancer Research* **40**, 1–19 (2021).
15. Batra, R., Manchanda, M. & Swanson, M. S. Global insights into alternative polyadenylation regulation. *RNA Biol* **12**, 597–602 (2015).
16. Masamha, C. P. *et al.* CFI_{m25} links alternative polyadenylation to glioblastoma tumour suppression. *Nature* **510**:7505 **510**, 412–416 (2014).
17. Martin, G., Gruber, A. R., Keller, W. & Zavolan, M. Genome-wide Analysis of Pre-mRNA 3' End Processing Reveals a Decisive Role of Human Cleavage Factor I in the Regulation of 3' UTR Length. *Cell Rep* **1**, 753–763 (2012).
18. Lutz, C. S. Alternative polyadenylation: A twist on mRNA 3' end formation. *ACS Chem Biol* **3**, 609–617 (2008).

19. Kwon, S. C. *et al.* The RNA-binding protein repertoire of embryonic stem cells. *Nat Struct Mol Biol* **20**, 1122–1130 (2013).
20. Berkovits, B. D. & Mayr, C. Alternative 3' UTRs act as scaffolds to regulate membrane protein localization. *Nature* **522**, 363–367 (2015).
21. D'Angeli, V. *et al.* Polypyrimidine tract binding protein 1 regulates the activation of mouse CD8 T cells. *Eur J Immunol* **52**, 1058 (2022).
22. Domingues, R. G. *et al.* CD5 expression is regulated during human T-cell activation by alternative polyadenylation, PTBP1 and miR-204. *Eur J Immunol* **46**, 1490 (2016).
23. Mayr, C. Regulation by 3'-Untranslated Regions. *Annu Rev Genet* **51**, 171–194 (2017).
24. García-Mauriño, S. M. *et al.* RNA binding protein regulation and cross-talk in the control of AU-rich mRNA Fate. *Front Mol Biosci* **4**, 297607 (2017).
25. Chen, C. A. & Shyu, A. Mechanisms of deadenylation-dependent decay. *WIREs RNA* **2**, 167–183 (2011).
26. Mattioli, C. C. *et al.* Alternative 3' UTRs direct localization of functionally diverse protein isoforms in neuronal compartments. *Nucleic Acids Res* **47**, 2560–2573 (2019).
27. Gaidatzis, D., van Nimwegen, E., Hausser, J. & Zavolan, M. Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics* **8**, 69 (2007).
28. Bartel, D. P. MicroRNAs: Target Recognition and Regulatory Functions. *Cell* **136**, 215–233 (2009).

29. Sanfilippo, P., Wen, J. & Lai, E. C. Landscape and evolution of tissue-specific alternative polyadenylation across *Drosophila* species. *Genome Biol* **18**, 1–22 (2017).
30. Smibert, P. *et al.* Global Patterns of Tissue-Specific Alternative Polyadenylation in *Drosophila*. *Cell Rep* **1**, 277–289 (2012).
31. Hong, W. *et al.* APAAtlas: decoding alternative polyadenylation across human tissues. *Nucleic Acids Res* **48**, D34–D39 (2020).
32. Miura, P., Shenker, S., Andreu-Agullo, C., Westholm, J. O. & Lai, E. C. Widespread and extensive lengthening of 3' UTRs in the mammalian brain. *Genome Res* **23**, 812–825 (2013).
33. Cao, J. & Kuyumcu-Martinez, M. N. Alternative polyadenylation regulation in cardiac development and cardiovascular disease. *Cardiovasc Res* **119**, 1324 (2023).
34. Reyes, A. *et al.* Drift and conservation of differential exon usage across tissues in primate species. *Proceedings of the National Academy of Sciences* **110**, 15377–15382 (2013).
35. Derti, A. *et al.* A quantitative atlas of polyadenylation in five mammals. *Genome Res* **22**, 1173–1183 (2012).
36. Xu, C. & Zhang, J. Alternative Polyadenylation of Mammalian Transcripts Is Generally Deleterious, Not Adaptive. *Cell Syst* **6**, 734-742.e4 (2018).
37. Lianoglou, S., Garg, V., Yang, J. L., Leslie, C. S. & Mayr, C. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev* **27**, 2380 (2013).

38. Shulman, E. D. & Elkon, R. Cell-type-specific analysis of alternative polyadenylation using single-cell transcriptomics data. *Nucleic Acids Res* **47**, 10027–10039 (2019).
39. Lee, S. *et al.* Diverse cell-specific patterns of alternative polyadenylation in *Drosophila*. *Nat Commun* **13**, 5372 (2022).
40. Zheng, D. *et al.* Cellular stress alters 3'UTR landscape through alternative polyadenylation and isoform-specific degradation. *Nat Commun* **9**, 2268 (2018).
41. Shepard, P. J. *et al.* Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA* **17**, 761–772 (2011).
42. Sandberg, R., Neilson, J. R., Sarma, A., Sharp, P. A. & Burge, C. B. Proliferating Cells Express mRNAs with Shortened 3' Untranslated Regions and Fewer MicroRNA Target Sites. *Science (1979)* **320**, 1643–1647 (2008).
43. Ji, Z., Lee, J. Y., Pan, Z., Jiang, B. & Tian, B. Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proceedings of the National Academy of Sciences* **106**, 7028–7033 (2009).
44. de Prisco, N. *et al.* Alternative polyadenylation alters protein dosage by switching between intronic and 3'UTR sites. *Sci Adv* **9**, (2023).
45. Wang, Q. *et al.* Cell Cycle Regulation by Alternative Polyadenylation of CCND1. *Scientific Reports 2018 8:1* **8**, 1–9 (2018).
46. Gallicchio, L., Olivares, G. H., Berry, C. W. & Fuller, M. T. Regulation and function of alternative polyadenylation in development and differentiation. *RNA Biol* **20**, 908 (2023).

47. Li, W. *et al.* Alternative cleavage and polyadenylation in spermatogenesis connects chromatin regulation with post-transcriptional control. *BMC Biol* **14**, 1–17 (2016).
48. Chang, J. W., Yeh, H. S. & Yong, J. Alternative Polyadenylation in Human Diseases. *Endocrinology and Metabolism* **32**, 413 (2017).
49. Curinha, A., Braz, S. O., Pereira-Castro, I., Cruz, A. & Moreira, A. Implications of polyadenylation in health and disease. *Nucleus* **5**, 508 (2014).
50. Mayr, C. & Bartel, D. P. Widespread Shortening of 3'UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells. *Cell* **138**, 673–684 (2009).
51. Lee, S.-H. *et al.* Widespread intronic polyadenylation inactivates tumour suppressor genes in leukaemia. *Nature* **561**, 127–131 (2018).
52. Xiao, S. *et al.* Relationship between NUDT21 mediated alternative polyadenylation process and tumor. *Front Oncol* **13**, (2023).
53. Xia, Z. *et al.* Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types. *Nat Commun* **5**, (2014).
54. Bacchetta, R., Barzaghi, F. & Roncarolo, M. G. From IPEX syndrome to FOXP3 mutation: A lesson on immune dysregulation. *Ann N Y Acad Sci* **1417**, 5–22 (2016).
55. Patel, R., Brophy, C., Hickling, M., Neve, J. & Furger, A. Alternative cleavage and polyadenylation of genes associated with protein turnover and mitochondrial function are deregulated in Parkinson's, Alzheimer's and ALS disease. *BMC Med Genomics* **12**, 60 (2019).

56. Romo, L., Ashar-Patel, A., Pfister, E. & Aronin, N. Alterations in mRNA 3' UTR Isoform Abundance Accompany Gene Expression Changes in Human Huntington's Disease Brains. *Cell Rep* **20**, 3057–3070 (2017).
57. Rhinn, H. *et al.* Alternative α -synuclein transcript usage as a convergent mechanism in Parkinson's disease pathology. *Nature Communications* **2012 3:1 3**, 1–11 (2012).
58. Ozawa, T., Kondo, M. & Isobe, M. 3' Rapid amplification of cDNA ends (RACE) walking for rapid structural analysis of large transcripts. *J Hum Genet* **49**, 102–105 (2004).
59. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463 (1977).
60. Bhardwaj, A. R., Pandey, R., Agarwal, M. & Katiyar-Agarwal, S. Northern Blotting Technique for Detection and Expression Analysis of mRNAs and Small RNAs. *Methods in Molecular Biology* **2170**, 155–183 (2021).
61. Zhang, H., Lee, J. & Tian, B. Biased alternative polyadenylation in human tissues. *Genome Biol* **6**, R100 (2005).
62. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nature Reviews Genetics* **2019 20:11 20**, 631–656 (2019).
63. Emrich, S. J., Barbazuk, W. B., Li, L. & Schnable, P. S. Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Res* **17**, 69–73 (2007).
64. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57 (2009).

65. Ye, C., Long, Y., Ji, G., Li, Q. Q. & Wu, X. APATrap: identification and quantification of alternative polyadenylation sites from RNA-seq data. *Bioinformatics* **34**, 1841–1849 (2018).
66. Ha, K. C. H., Blencowe, B. J. & Morris, Q. QAPA: A new method for the systematic analysis of alternative polyadenylation from RNA-seq data. *Genome Biol* **19**, 1–18 (2018).
67. Zheng, D., Liu, X. & Tian, B. 3' READS+, a sensitive and accurate method for 3' end sequencing of polyadenylated RNA. (2016) doi:10.1261/rna.057075.116.
68. Corley, S. M., Troy, N. M., Bosco, A. & Wilkins, M. R. QuantSeq. 3' Sequencing combined with Salmon provides a fast, reliable approach for high throughput RNA expression analysis. *Sci Rep* **9**, 1–15 (2019).
69. Ogorodnikov, A. & Danckwardt, S. TRENDseq—A highly multiplexed high throughput RNA 3' end sequencing for mapping alternative polyadenylation. *Methods Enzymol* **655**, 37–72 (2021).
70. Chen, Y. *et al.* A systematic benchmark of Nanopore long read RNA sequencing for transcript level analysis in human cell lines. *bioRxiv* 2021.04.21.440736 (2021) doi:10.1101/2021.04.21.440736.
71. Joglekar, A. *et al.* Single-cell long-read sequencing-based mapping reveals specialized splicing patterns in developing and adult mouse and human brain. *Nat Neurosci* **27**, 1051–1063 (2024).
72. Workman, R. E. *et al.* Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nature Methods* 2019 16:12 **16**, 1297–1305 (2019).

73. Wang, B. *et al.* Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nature Communications* 2016 7:1 **7**, 1–13 (2016).
74. Wohlers, I., Garg, S. & Hehir-Kwa, J. Y. Editorial: Long-read sequencing—Pitfalls, benefits and success stories. *Front Genet* **13**, 1114542 (2023).
75. Amarasinghe, S. L. *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biology* 2020 21:1 **21**, 1–16 (2020).
76. Ziegenhain, C. *et al.* Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol Cell* **65**, 631-643.e4 (2017).
77. Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics* 2015 16:3 **16**, 133–145 (2015).
78. Conte, M. I., Fuentes-Trillo, A. & Domínguez Conde, C. Opportunities and tradeoffs in single-cell transcriptomic technologies. *Trends in Genetics* **40**, 83–93 (2024).
79. Fansler, M. M., Mitschka, S. & Mayr, C. Quantifying 3'UTR length from scRNA-seq data reveals changes independent of gene expression. *Nature Communications* 2024 15:1 **15**, 1–18 (2024).
80. Gao, Y., Li, L., Amos, C. I. & Li, W. Analysis of alternative polyadenylation from single-cell RNA-seq using scDaPars reveals cell subpopulations invisible to gene expression. *Genome Res* **31**, 1856–1866 (2021).
81. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* **6**, 377–382 (2009).
82. Islam, S. *et al.* Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res* **21**, 1160–1167 (2011).

83. Jaitin, D. A. *et al.* Massively parallel single cell RNA-Seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776 (2014).
84. Hochgerner, H. *et al.* STRT-seq-2i: dual-index 5' single cell and nucleus RNA-seq on an addressable microwell array. *Scientific Reports* 2017 7:1 **7**, 1–8 (2017).
85. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell Rep* **2**, 666–673 (2012).
86. Ramsköld, D. *et al.* Full-Length mRNA-Seq from single cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol* **30**, 777 (2012).
87. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols* 2013 9:1 **9**, 171–181 (2014).
88. Fan, H. C., Fu, G. K. & Fodor, S. P. A. Combinatorial labeling of single cells for gene expression cytometry. *Science (1979)* **347**, (2015).
89. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols* 2018 13:4 **13**, 599–604 (2018).
90. Rosenberg, A. B. *et al.* Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science (1979)* **360**, 176–182 (2018).
91. Cao, J. *et al.* Comprehensive single cell transcriptional profiling of a multicellular organism. *Science* **357**, 661 (2017).
92. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* 2019 566:7745 **566**, 496–502 (2019).
93. Yin, H. & Marshall, D. Microfluidics for single cell analysis. *Curr Opin Biotechnol* **23**, 110–119 (2012).

94. DeLaughter, D. M. The use of the Fluidigm C1 for RNA expression analyses of single cells. *Curr Protoc Mol Biol* **122**, e55 (2018).
95. Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
96. Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
97. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature Communications* 2017 8:1 **8**, 1–12 (2017).
98. Lafzi, A., Moutinho, C., Picelli, S. & Heyn, H. Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies. *Nat Protoc* **13**, 2742–2757 (2018).
99. Alles, J. *et al.* Cell fixation and preservation for droplet-based single-cell transcriptomics. *BMC Biol* **15**, 1–14 (2017).
100. Wang, W., Penland, L., Gokce, O., Croote, D. & Quake, S. R. High fidelity hypothermic preservation of primary tissues in organ transplant preservative for single cell transcriptome analysis. *BMC Genomics* **19**, 1–10 (2018).
101. Guillaumet-Adkins, A. *et al.* Single-cell transcriptome conservation in cryopreserved cells and tissues. *Genome Biol* **18**, (2017).
102. Habib, N. *et al.* Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nature Methods* 2017 14:10 **14**, 955–958 (2017).
103. Habib, N. *et al.* Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. *Science* (1979) **353**, 925–928 (2016).
104. Paul, F. *et al.* Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. *Cell* **163**, 1663–1677 (2015).

105. Prakadan, S. M., Shalek, A. K. & Weitz, D. A. Scaling by shrinking: empowering single-cell 'omics' with microfluidic devices. *Nature Reviews Genetics* 2017 18:6 **18**, 345–361 (2017).
106. Mardis, E. R. Next-generation sequencing platforms. *Annual Review of Analytical Chemistry* **6**, 287–303 (2013).
107. Fox, E. J., Reid-Bayliss, K. S., Emond, M. J. & Loeb, L. A. Accuracy of Next Generation Sequencing Platforms. *Next Gener Seq Appl* **1**, (2014).
108. Svensson, V. *et al.* Power analysis of single-cell RNA-sequencing experiments. *Nature Methods* 2017 14:4 **14**, 381–387 (2017).
109. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol* **37**, 38–44 (2019).
110. Kobak, D. & Berens, P. The art of using t-SNE for single-cell transcriptomics. *Nature Communications* 2019 10:1 **10**, 1–14 (2019).
111. Kiselev, V. Y., Andrews, T. S. & Hemberg, M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics* 2018 20:5 **20**, 273–282 (2019).
112. Feng, C. *et al.* Dimension Reduction and Clustering Models for Single-Cell RNA Sequencing Data: A Comparative Study. *Int J Mol Sci* **21**, 2181 (2020).
113. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
114. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381–386 (2014).

115. Hedlund, E. & Deng, Q. Single-cell RNA sequencing: Technical advancements and biological applications. *Mol Aspects Med* **59**, 36–46 (2018).
116. Hagemann-Jensen, M. *et al.* Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nature Biotechnology* 2020 38:6 **38**, 708–714 (2020).
117. Fan, X. *et al.* Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. *Genome Biol* **16**, 1–17 (2015).
118. Sheng, K., Cao, W., Niu, Y., Deng, Q. & Zong, C. Effective detection of variation in single-cell transcriptomes using MATQ-seq. *Nat Methods* **14**, 267–270 (2017).
119. Mora-Castilla, S. *et al.* Miniaturization Technologies for Efficient Single-Cell Library Preparation for Next-Generation Sequencing. *SLAS Technol* **21**, 557–567 (2016).
120. Jaeger, B. N. *et al.* Miniaturization of Smart-seq2 for Single-Cell and Single-Nucleus RNA Sequencing. *STAR Protoc* **1**, 100081 (2020).
121. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
122. Merker, J. D. *et al.* Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genetics in Medicine* **20**, 159–163 (2018).
123. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology* 2016 17:1 **17**, 1–11 (2016).

124. Tardaguila, M. *et al.* SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res* **28**, 396–411 (2018).
125. Jain, M. *et al.* Improved data analysis for the MinION nanopore sequencer. *Nature Methods* 2015 12:4 **12**, 351–356 (2015).
126. Jaworski, E. & Routh, A. Parallel ClickSeq and Nanopore sequencing elucidates the rapid evolution of defective-interfering RNAs in Flock House virus. *PLoS Pathog* **13**, e1006365 (2017).
127. Phipson, B. *et al.* Gene length and detection bias in single cell RNA sequencing protocols. *F1000Research* 2017 6:595 **6**, 595 (2017).
128. Saliba, A. E., Westermann, A. J., Gorski, S. A. & Vogel, J. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res* **42**, 8845 (2014).
129. Tran, H. T. N. *et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol* **21**, 1–32 (2020).
130. 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).
131. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* **16**, 1289–1296 (2019).
132. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).
133. Shaham, U. *et al.* Removal of batch effects using distribution-matching residual networks. *Bioinformatics* **33**, 2539–2546 (2017).
134. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol* **15**, e8746 (2019).

135. Denisenko, E. *et al.* Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. *Genome Biol* **21**, 1–25 (2020).
136. García-Castro, H. *et al.* ACME dissociation: a versatile cell fixation-dissociation method for single-cell transcriptomics. *Genome Biol* **22**, 89 (2021).
137. Liu, L. *et al.* Dissociation of microdissected mouse brain tissue for artifact free single-cell RNA sequencing. *STAR Protoc* **2**, 100590 (2021).
138. Bresciani, E., Broadbridge, E. & Liu, P. P. An efficient dissociation protocol for generation of single cell suspension from zebrafish embryos and larvae. *MethodsX* **5**, 1287–1290 (2018).
139. Van Den Brink, S. C. *et al.* Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nature Methods* **2017** 14:10 **14**, 935–936 (2017).
140. Howanitz, J. H. & Howanitz, P. J. Laboratory Results: Timeliness as a Quality Attribute and Strategy. *Am J Clin Pathol* **116**, 311–315 (2001).
141. Gallion, L. A., Anttila, M. M., Abraham, D. H., Proctor, A. & Allbritton, N. L. Preserving single cells in space and time for analytical assays. *TrAC - Trends in Analytical Chemistry* **122**, (2020).
142. Wohnhaas, C. T. *et al.* DMSO cryopreservation is the method of choice to preserve cells for droplet-based single-cell RNA sequencing. *Sci Rep* **9**, (2019).
143. Chen, J., Cheung, F., Shi, R., Zhou, H. & Lu, W. PBMC fixation and processing for Chromium single-cell RNA sequencing. *J Transl Med* **16**, 198 (2018).

144. Phan, H. Van *et al.* High-throughput RNA sequencing of paraformaldehyde-fixed single cells. *Nat Commun* **12**, 5636 (2021).
145. Attar, M. *et al.* A practical solution for preserving single cells for RNA sequencing. *Scientific Reports* 2018 8:1 **8**, 1–10 (2018).
146. Wang, X., Yu, L. & Wu, A. R. The effect of methanol fixation on single-cell RNA sequencing data. *BMC Genomics* **22**, 1–16 (2021).
147. Raddi, G. *et al.* Mosquito cellular immunity at single-cell resolution. *Science* (1979) **369**, 1128–1132 (2020).
148. Fortmann, S. D. *et al.* Fixation Before Dissociation Using a Deep Eutectic Solvent Preserves In Vivo States and Phospho-Signaling in Single-Cell Sequencing. *bioRxiv* 2023.02.13.528370 (2023) doi:10.1101/2023.02.13.528370.
149. Ye, W., Lian, Q., Ye, C. & Wu, X. A Survey on Methods for Predicting Polyadenylation Sites from DNA Sequences, Bulk RNA-seq, and Single-cell RNA-seq. *Genomics Proteomics Bioinformatics* **21**, 67–83 (2023).
150. Patrick, R. *et al.* Sierra: Discovery of differential transcript usage from polyA-captured single-cell RNA-seq data. *Genome Biol* **21**, 1–27 (2020).
151. Wu, X., Liu, T., Ye, C., Ye, W. & Ji, G. scAPATrap: identification and quantification of alternative polyadenylation sites from single-cell RNA-seq data. *Brief Bioinform* **22**, (2021).
152. Li, G. W. *et al.* SCAPTURE: a deep learning-embedded pipeline that captures polyadenylation information from 3' tag-based RNA-seq of single cells. *Genome Biol* **22**, 1–24 (2021).
153. Ye, C., Lin, J. & Li, Q. Q. Discovery of alternative polyadenylation dynamics from single cell types. *Comput Struct Biotechnol J* **18**, 1012 (2020).

154. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576 (2010).
155. Li, W. V., Zheng, D., Wang, R. & Tian, B. MAAPER: model-based analysis of alternative polyadenylation using 3' end-linked reads. *Genome Biol* **22**, 1–21 (2021).
156. Melsted, P. *et al.* Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nature Biotechnology* 2021 39:7 **39**, 813–818 (2021).
157. Jerber, J. *et al.* Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation. *Nature Genetics* 2021 53:3 **53**, 304–312 (2021).
158. Shi, Y., Kirwan, P. & Livesey, F. J. Directed differentiation of human pluripotent stem cells to cerebral cortex neurons and neural networks. *Nature Protocols* 2012 7:10 **7**, 1836–1846 (2012).
159. Broad Institute. Java tools for analyzing Drop-seq data. Preprint at (2019).
160. Howe, K. L. *et al.* Ensembl Genomes 2020—enabling non-vertebrate genomic research. *Nucleic Acids Res* **48**, D689–D695 (2020).
161. Broad Institute. Picard toolkit. *Broad Institue, GitHub repository* (2018).
162. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).
163. McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst* **8**, 329-337.e4 (2019).
164. Zappia, L. & Oshlack, A. Clustering trees: a visualization for evaluating clusterings at multiple resolutions. *Gigascience* **7**, (2018).

165. Büttner, M., Ostner, J., Müller, C. L., Theis, F. J. & Schubert, B. scCODA is a Bayesian model for compositional single-cell data analysis. *Nat Commun* **12**, 6876 (2021).
166. Crowell, H. L. *et al.* muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nature Communications* 2020 11:1 **11**, 1–12 (2020).
167. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
168. Fang, Z., Liu, X. & Peltz, G. GSEAPy: a comprehensive package for performing gene set enrichment analysis in Python. *Bioinformatics* **39**, (2023).
169. Raivo Kolde. pheatmap: Pretty Heatmaps. Preprint at (2018).
170. Kimes, P. K., Liu, Y., Neil Hayes, D. & Marron, J. S. Statistical Significance for Hierarchical Clustering. *Biometrics* **73**, 811–821 (2017).
171. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47**, D766–D773 (2019).
172. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**, 417–419 (2017).
173. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (2021).
174. Neph, S. *et al.* BEDOPS: high-performance genomic feature operations. *Bioinformatics* **28**, 1919–1920 (2012).

175. Lawrence, M. *et al.* Software for Computing and Annotating Genomic Ranges. *PLoS Comput Biol* **9**, e1003118 (2013).
176. Hahne, F. & Ivanek, R. Visualizing genomic data using Gviz and bioconductor. *Methods in Molecular Biology* **1418**, 335–351 (2016).
177. Zappia, L., Phipson, B. & Oshlack, A. Splatter: Simulation of single-cell RNA sequencing data. *Genome Biol* **18**, 1–15 (2017).
178. Lukassen, S., Bosch, E., Ekici, A. B. & Winterpacht, A. Single-cell RNA sequencing of adult mouse testes. *Sci Data* **5**, (2018).
179. Church, D. M. *et al.* Modernizing Reference Genome Assemblies. *PLoS Biol* **9**, e1001091 (2011).
180. Hao, Y. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol* **42**, 293–304 (2024).
181. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
182. Frankish, A. *et al.* GENCODE 2021. *Nucleic Acids Res* **49**, D916–D923 (2021).
183. R Core Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing* (2021).
184. Feng, X., Li, L., Wagner, E. J. & Li, W. TC3A: The Cancer 3' UTR Atlas. *Nucleic Acids Res* **46**, D1027 (2018).
185. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938–2940 (2017).

186. Diag, A., Schilling, M., Klironomos, F., Ayoub, S. & Rajewsky, N. Spatiotemporal m(i)RNA Architecture and 3' UTR Regulation in the *C. elegans* Germline. *Dev Cell* **47**, 785-800.e8 (2018).
187. Green, M. R. & Sambrook, J. *Molecular Cloning: A Laboratory Manual*. (Cold Spring Harbor Laboratory Press, 2012).
188. Scotto-Lavino, E., Du, G. & Frohman, M. A. 3' end cDNA amplification using classic RACE. *Nat Protoc* **1**, 2742–2745 (2006).
189. Arancio, W., Sciaraffa, N. & Coronello, C. MBS: a genome browser annotation track for high-confident microRNA binding sites in whole human transcriptome. *Database* **2023**, 1–7 (2023).
190. Kozomara, A., Birgaoanu, M. & Griffiths-Jones, S. miRBase: from microRNA sequences to function. *Nucleic Acids Res* **47**, D155–D162 (2019).
191. Tsujimura, K., Shiohama, T. & Takahashi, E. microRNA Biology on Brain Development and Neuroimaging Approach. *Brain Sci* **12**, (2022).
192. Lukassen, S., Bosch, E., Ekici, A. B. & Winterpacht, A. Characterization of germ cell differentiation in the male mouse through single-cell RNA sequencing. *Scientific Reports* **2018 8:1** **8**, 1–7 (2018).
193. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
194. Xie, Z. *et al.* Gene Set Knowledge Discovery with Enrichr. *Curr Protoc* **1**, e90 (2021).
195. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat Genet* **25**, 25–29 (2000).

196. Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York (2016).
197. Zhang, X. *et al.* CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res* **47**, D721–D728 (2019).
198. Abugessaisa, I., Noguchi, S., Carninci, P. & Kasukawa, T. The FANTOM5 Computation Ecosystem: Genomic Information Hub for Promoters and Active Enhancers. *Methods Mol Biol* **1611**, 199–217 (2017).
199. Moon, Y., Herrmann, C. J., Mironov, A. & Zavolan, M. PolyASite v3.0: a multi-species atlas of polyadenylation sites inferred from single-cell RNA-sequencing data. *Nucleic Acids Res* **53**, D197–D204 (2025).
200. Frankish, A. *et al.* GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res* **51**, D942 (2022).
201. Gutiérrez-Franco, A. *et al.* Methanol fixation is the method of choice for droplet-based single-cell transcriptomics of neural cells. *Communications Biology* 2023 6:1 **6**, 1–12 (2023).
202. Illicic, T. *et al.* Classification of low quality cells from single-cell RNA-seq data. *Genome Biol* **17**, 1–15 (2016).
203. Nowakowski, T. J. *et al.* Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science* (1979) **358**, 1318–1323 (2017).
204. Bakken, T. E. *et al.* Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. *PLoS One* **13**, e0209648 (2018).
205. Ake, F. *et al.* Quantification of transcript isoforms at the single-cell level using SCALPEL. *Nature Communications* 2025 16:1 **16**, 1–17 (2025).

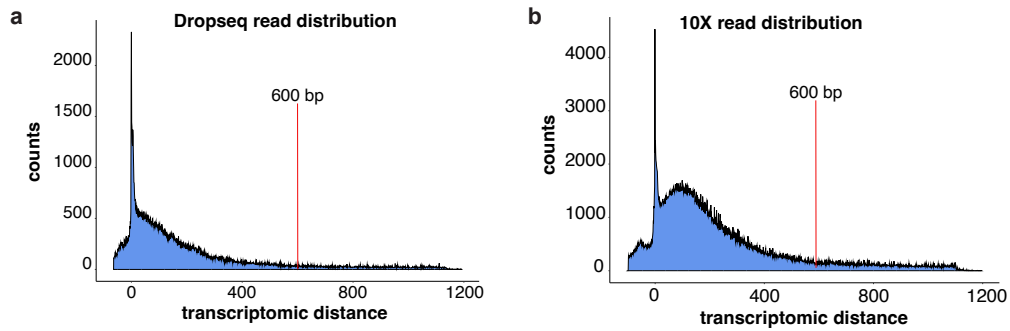
206. DI Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nature Biotechnology* 2017 35:4 **35**, 316–319 (2017).
207. Sarkar, H., Srivastava, A. & Patro, R. Minnow: a principled framework for rapid simulation of dscRNA-seq data at the read level. *Bioinformatics* **35**, i136–i144 (2019).
208. Liu, H. *et al.* IFT25, an intraflagellar transporter protein dispensable for ciliogenesis in somatic cells, is essential for sperm flagella formation†. *Biol Reprod* **96**, 993–1006 (2017).
209. Sironen, A. *et al.* Loss of SPEF2 Function in Mice Results in Spermatogenesis Defects and Primary Ciliary Dyskinesia1. *Biol Reprod* **85**, 690–701 (2011).
210. Yogo, K. Molecular basis of the morphogenesis of sperm head and tail in mice. *Reprod Med Biol* **21**, (2022).
211. Meng, G.-Q. *et al.* Bi-allelic variants in *DNAH3* cause male infertility with asthenoteratozoospermia in humans and mice. *Hum Reprod Open* **2024**, (2024).
212. Sutherland, J. M. *et al.* RNA binding protein Musashi-1 directly targets Msi2 and Erh during early testis germ cell development and interacts with IPO5 upon translocation to the nucleus. *The FASEB Journal* **29**, 2759–2768 (2015).
213. Vanderzwalmen, P. *et al.* The problems of spermatid microinjection in the human: the need for an accurate morphological approach and selective methods for viable and normal cells. *Human Reproduction* **13**, 515–519 (1998).

214. Ameri, A. J. & Lewis, Z. A. Shannon entropy as a metric for conditional gene expression in *Neurospora crassa*. *G3 Genes|Genomes|Genetics* **11**, (2021).
215. Liu, D. *et al.* Systematic variation in mRNA 3'-processing signals during mouse spermatogenesis. *Nucleic Acids Res* **35**, 234–246 (2007).
216. Guvenek, A. & Tian, B. Analysis of alternative cleavage and polyadenylation in mature and differentiating neurons using RNA-seq data. *Quantitative Biology* **6**, 253–266 (2018).
217. Kawahara, H., Imai, T. & Okano, H. MicroRNAs in neural stem cells and neurogenesis. *Front Neurosci* **6**, 1–13 (2012).
218. Lang, M. F. & Shi, Y. Dynamic roles of microRNAs in neurogenesis. *Front Neurosci* **6**, 22665 (2012).
219. Jonas, S. & Izaurralde, E. Towards a molecular understanding of microRNA-mediated gene silencing. *Nature Reviews Genetics* **2015** 16:7 **16**, 421–433 (2015).
220. Li, R. *et al.* Differentiation of Human iPS Cells Into Sensory Neurons Exhibits Developmental Stage-Specific Cryopreservation Challenges. *Front Cell Dev Biol* **9**, 796960 (2021).
221. Verrillo, L. *et al.* A reliable strategy for single-cell RNA sequencing analysis using cryoconserved primary cortical cells. *J Neurosci Methods* **347**, 108960 (2021).
222. Li, R., Hornberger, K., Dutton, J. R. & Hubel, A. Cryopreservation of Human iPS Cell Aggregates in a DMSO-Free Solution—An Optimization and Comparative Study. *Front Bioeng Biotechnol* **8**, (2020).

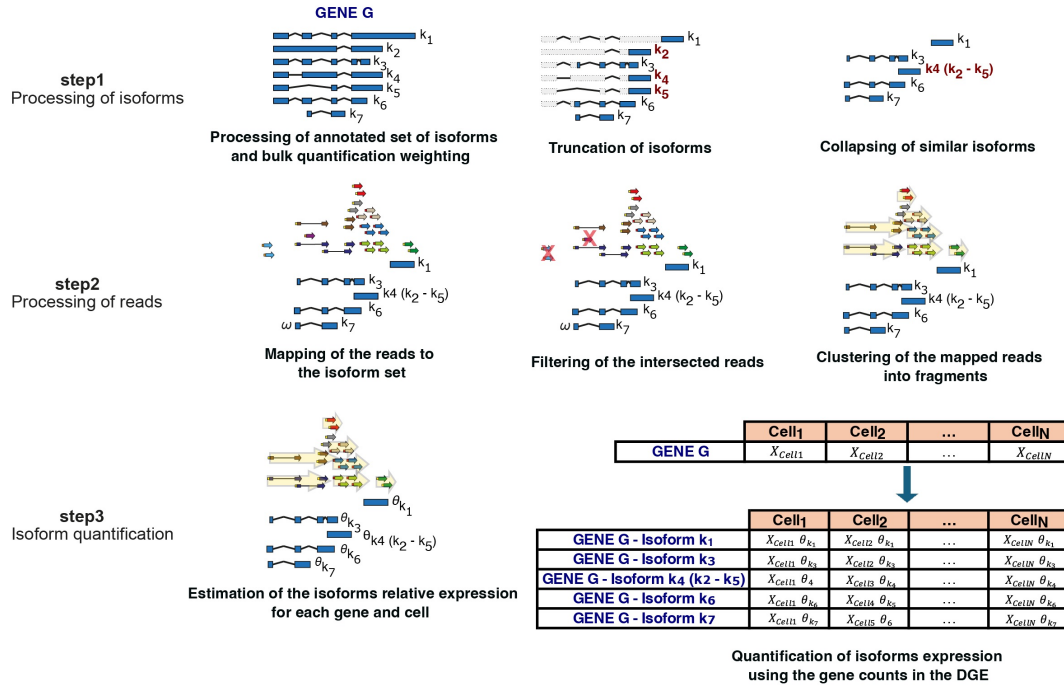
223. Niehaus, J. K., Taylor-Blake, B., Loo, L., Simon, J. M. & Zylka, M. J. Spinal macrophages resolve nociceptive hypersensitivity after peripheral injury. *Neuron* **109**, 1274-1282.e6 (2021).
224. Inak, G. *et al.* Defective metabolic programming impairs early neuronal morphogenesis in neural cultures and an organoid model of Leigh syndrome. *Nature Communications* 2021 12:1 **12**, 1–22 (2021).
225. Spišáková, V., Hejtmánek, L. & Hynšt, J. Nextflow in Bioinformatics: Executors Performance Comparison Using Genomics Data. *Future Generation Computer Systems* **142**, 328–339 (2023).
226. Langer, B. E. *et al.* Empowering bioinformatics communities with Nextflow and nf-core. *bioRxiv* 2024.05.10.592912 (2024) doi:10.1101/2024.05.10.592912.
227. Patowary, A. *et al.* Developmental isoform diversity in the human neocortex informs neuropsychiatric risk mechanisms. *Science (1979)* **384**, (2024).
228. Boeshaghi, A. S. *et al.* Isoform cell-type specificity in the mouse primary motor cortex. *Nature* **598**, 195–199 (2021).
229. Sommerkamp, P. *et al.* Differential Alternative Polyadenylation Landscapes Mediate Hematopoietic Stem Cell Activation and Regulate Glutamine Metabolism. *Cell Stem Cell* **26**, 722-738.e7 (2020).

8. Appendices

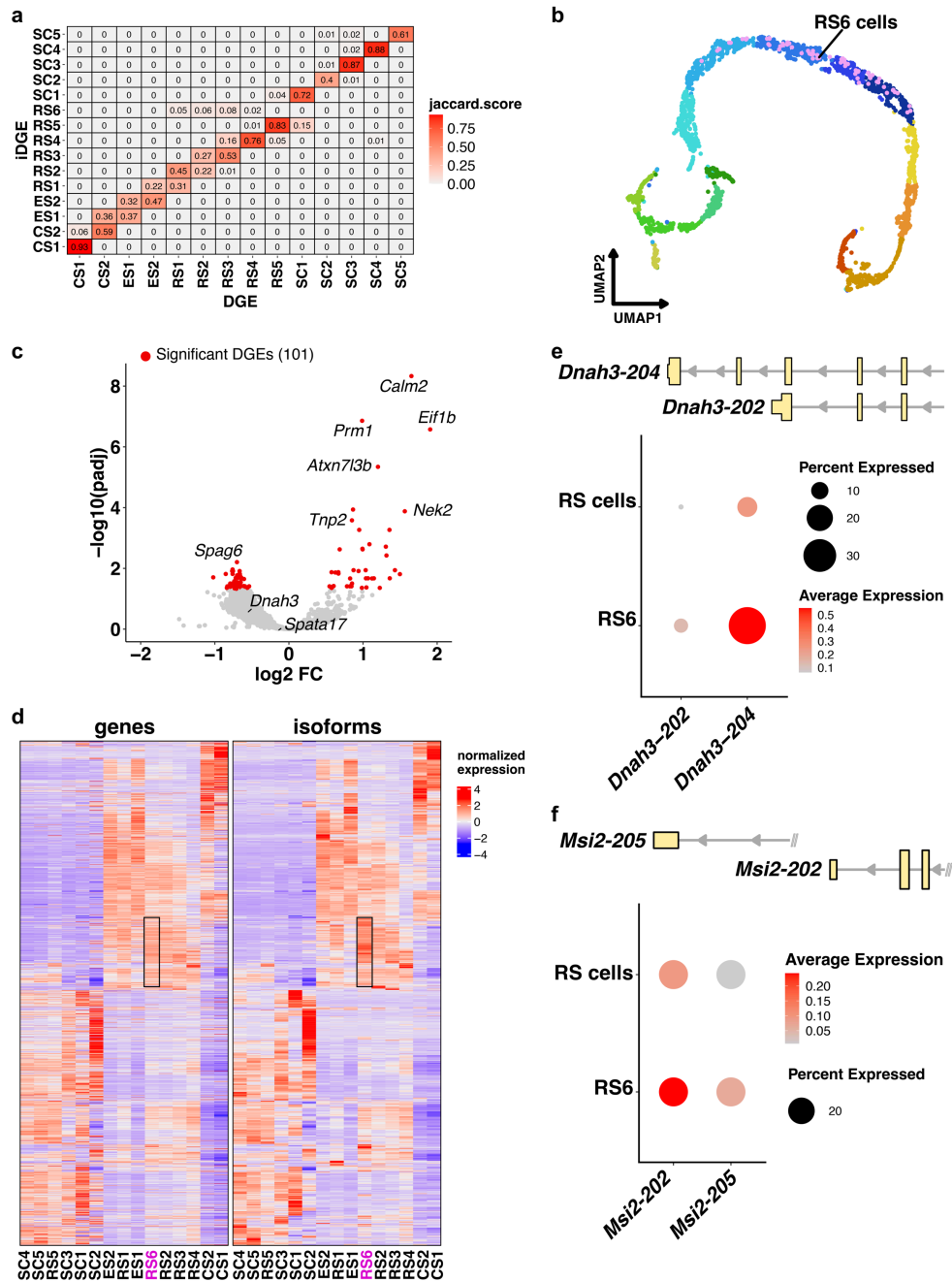
8.1. Supplementary Figures



Supplementary figure 1. Distance distribution of reads to annotated 3' ends of genes with only one expressed isoform. Empiric distribution of the 3' end of mapped reads in the transcriptomic space relative to isoform 3' ends in the drop-seq dataset **(a)** and the 10x dataset **(b)**. These distributions are used in SCALPEL to define the probability of a read to come from a particular isoform.

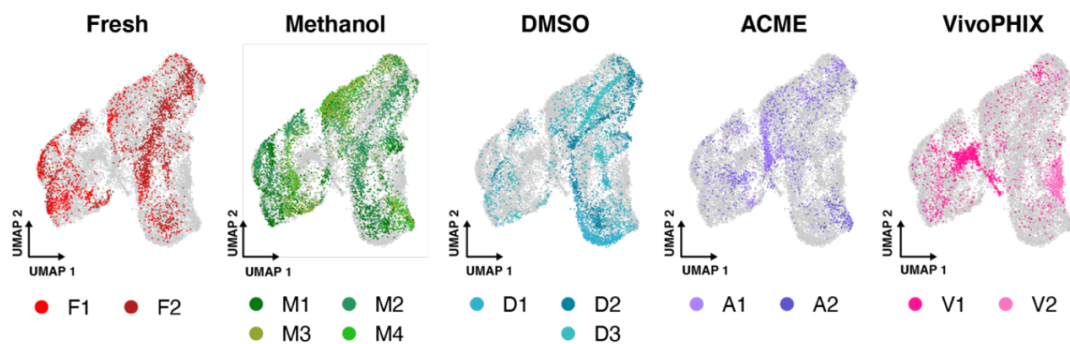


Supplementary figure 2. Detailed SCALPEL workflow. SCALPEL performs isoform quantification in 3 steps. Step 1. Isoforms are truncated to the last 600 nt before collapsing similar isoforms. Step 2. All the reads annotated to genes in the BAM file are mapped to the set of collapsed isoforms. Intronic or IP-derived reads are discarded. Reads with the same cell barcode and UMI are then assembled in fragments. 3. Relative isoform abundances are estimated for each cell using the fragment probabilities. Once isoform probabilities are estimated, final isoform expression is calculated by multiplying the relative isoform abundances by the UMI counts assigned to each gene in each cell.

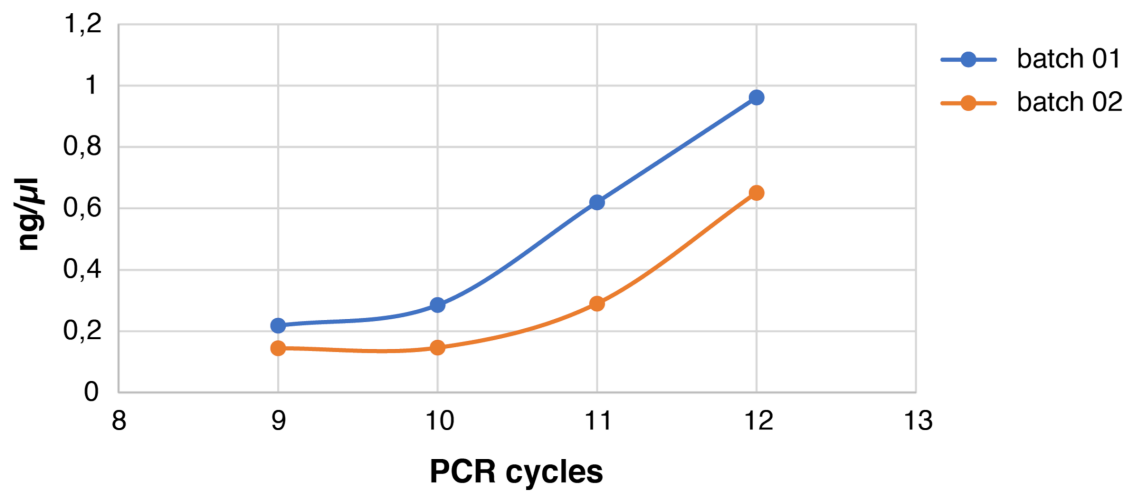


Supplementary figure 3. Differences between gene and isoform-based clustering solutions on the mouse spermatogenesis dataset. **a.** Similarity in the cell classification across different clusters between the isoform- (iDGE) and the gene- (DGE) based analyses. All the clusters in the iDGE analysis have a one-to-one correspondence with DGE clusters except for RS6 cluster. **b.** Cells from the RS6 cluster are highlighted on the UMAP generated using gene expression data. In this plot, RS6 cells are spread across multiple RS clusters (RS2-RS5) and are not located within any specific cluster. **c.** Differential gene expression analysis between RS6 cells and other RS cells identifies few differentially expressed genes. **d.** Heatmap showing

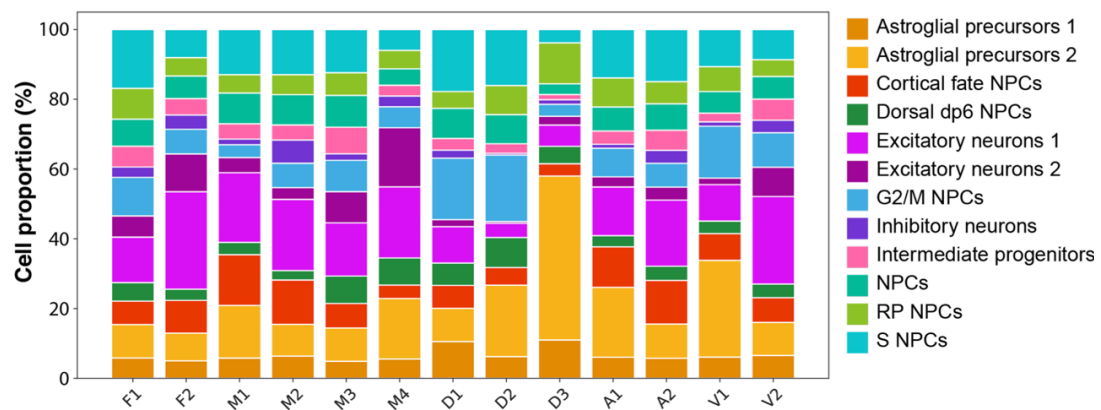
the expression of cluster markers identified in the isoform-based analysis (right) and their corresponding genes (left). The expression of some of the top RS6 isoform markers (black box, right) is not recapitulated at the gene level. Isoform expression distinguishes clusters more clearly than gene expression. **e, f.** DotPlot of RS6 marker isoforms of *Dnah3* (**d**) and *Msi2* (**e**) genes across RS cell populations. The plots highlight the differences in the expression of these isoforms in RS6 cells compared to other RS cells.



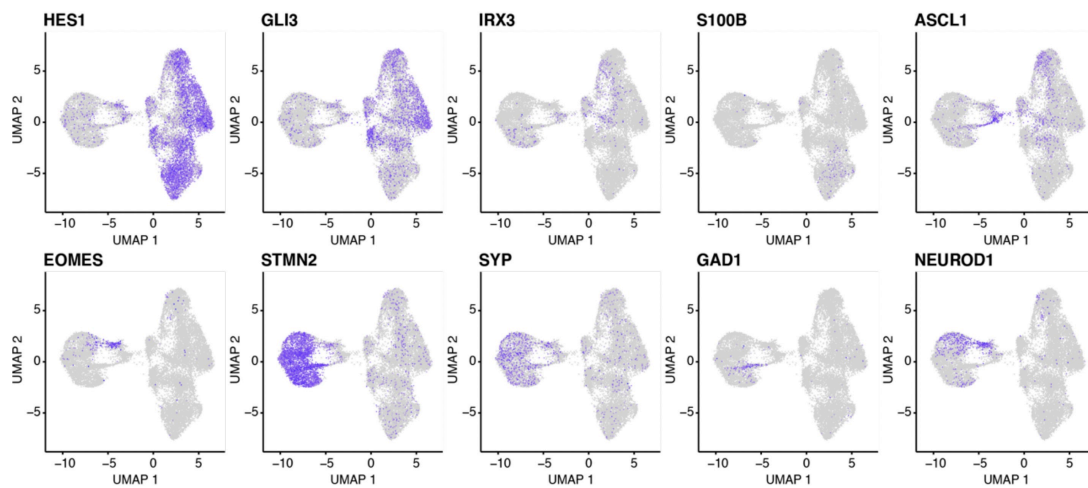
Supplementary figure 4. Cell preservation induces biases in gene expression that affect sample comparison. UMAP plots showing the distribution of cells from each sample on a UMAP before sample integration. The samples for each processing method (Fresh, Methanol, DMSO, ACME and vivoPHIX) are shown in separate plots. As can be seen from the analyses, cells processed using different methods show different distributions on the UMAP, which highlights the presence of fixation-specific batch effects.



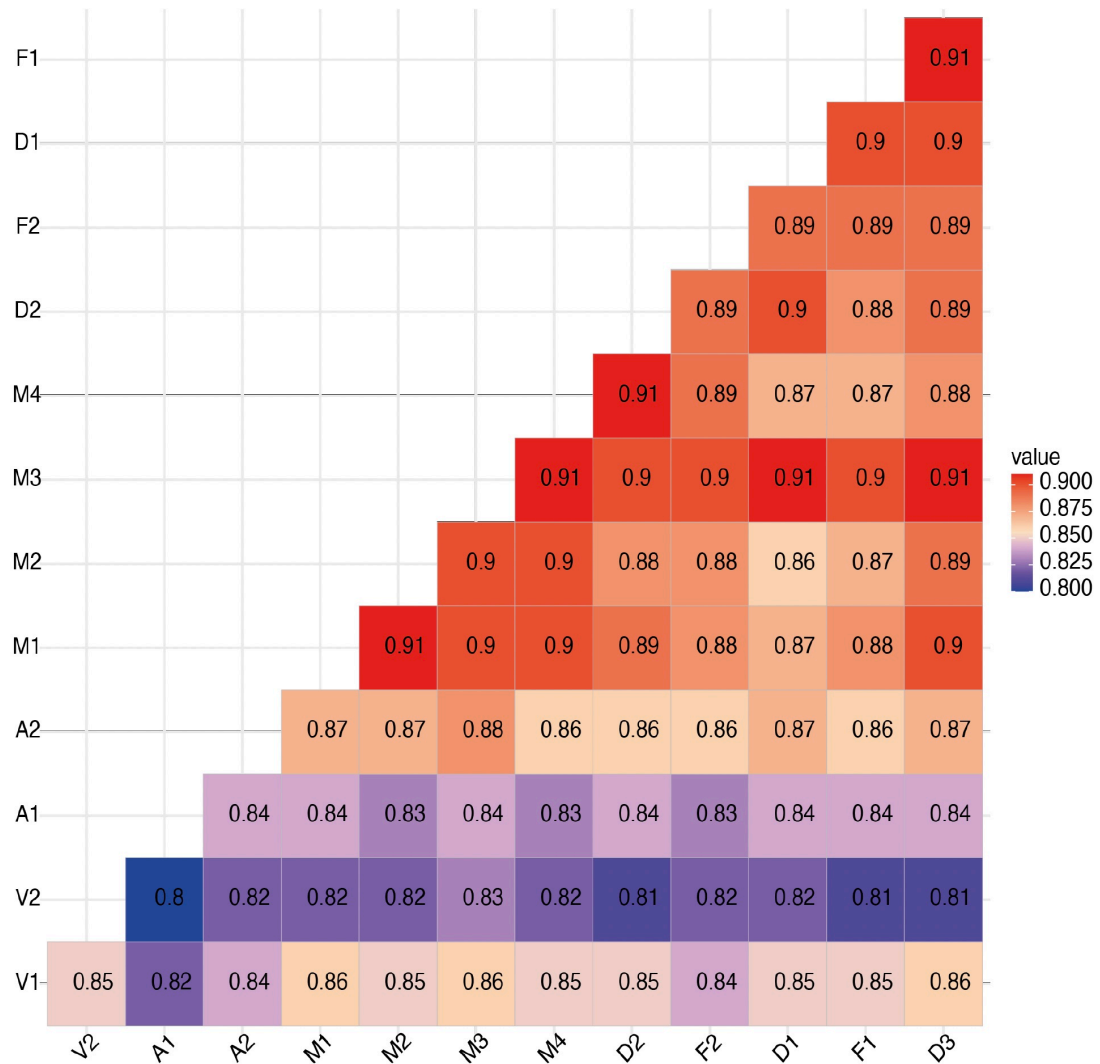
Supplementary figure 5. Comparison of the capture efficiency of the 2 bead batches used in the project. The plots show the concentration of cDNA in ng μl^{-1} as a function of PCR cycles. Beads from batch 01 have a higher capture efficiency than batch 02, which is reflected in a higher number of genes and UMIs in the samples processed with these beads



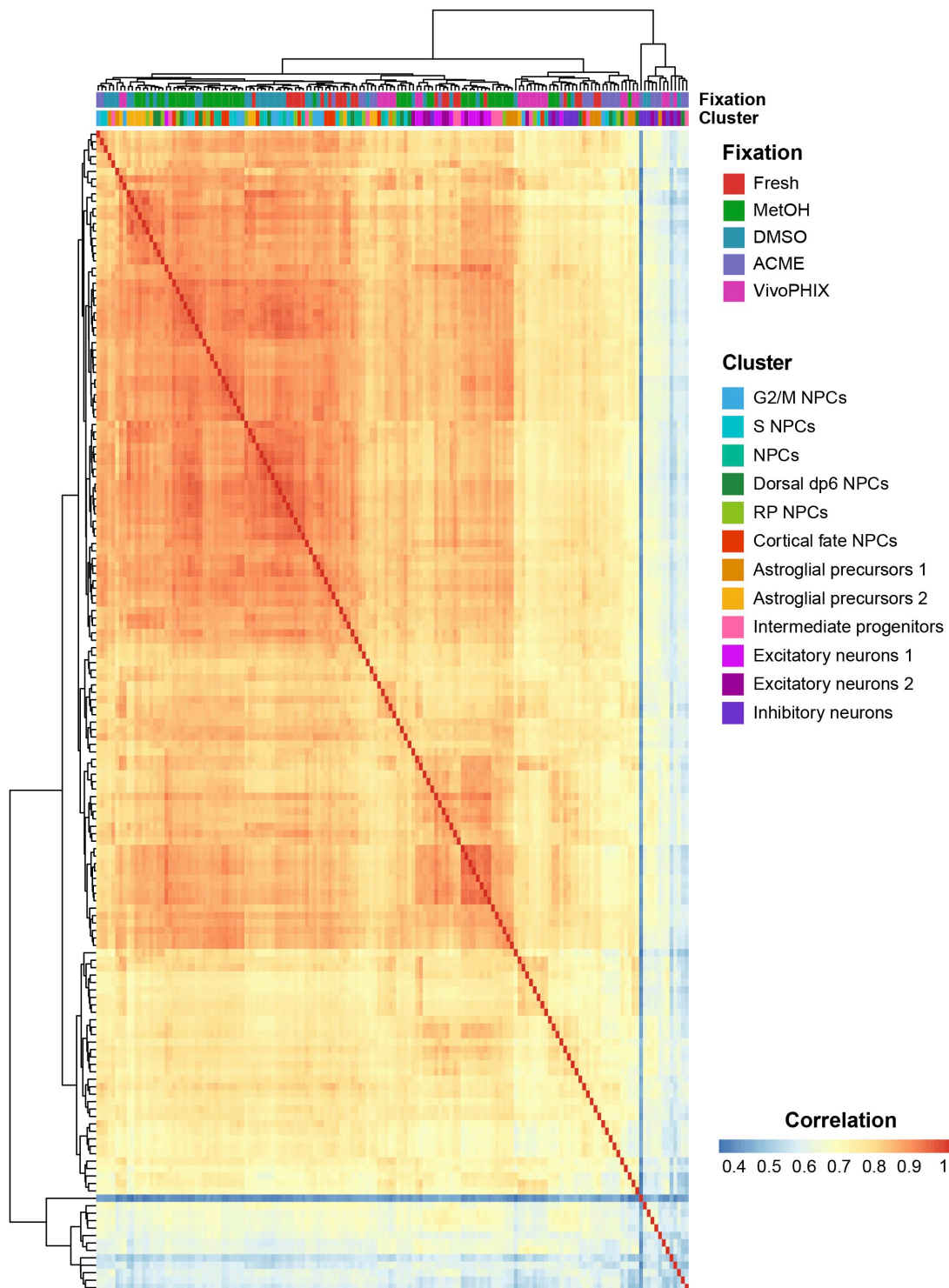
Supplementary figure 6. Cell composition biases across samples. Stacked barplots showing the proportion of cells in each of the samples. Cell composition changes across samples depends on the cell lines used, the differentiation experiment and the fixation/preservation method used. The analysis with scCODA only identified a significant depletion in excitatory neuron populations in DMSO cryopreserved samples (light and dark purple boxes). The samples are labelled as follows. fresh (F1 and F2), methanol (M1, M2, M3 and M4), DMSO (D1, D2 and D3), ACME (A1 and A2,) and vivoPHIX (V1 and V2).



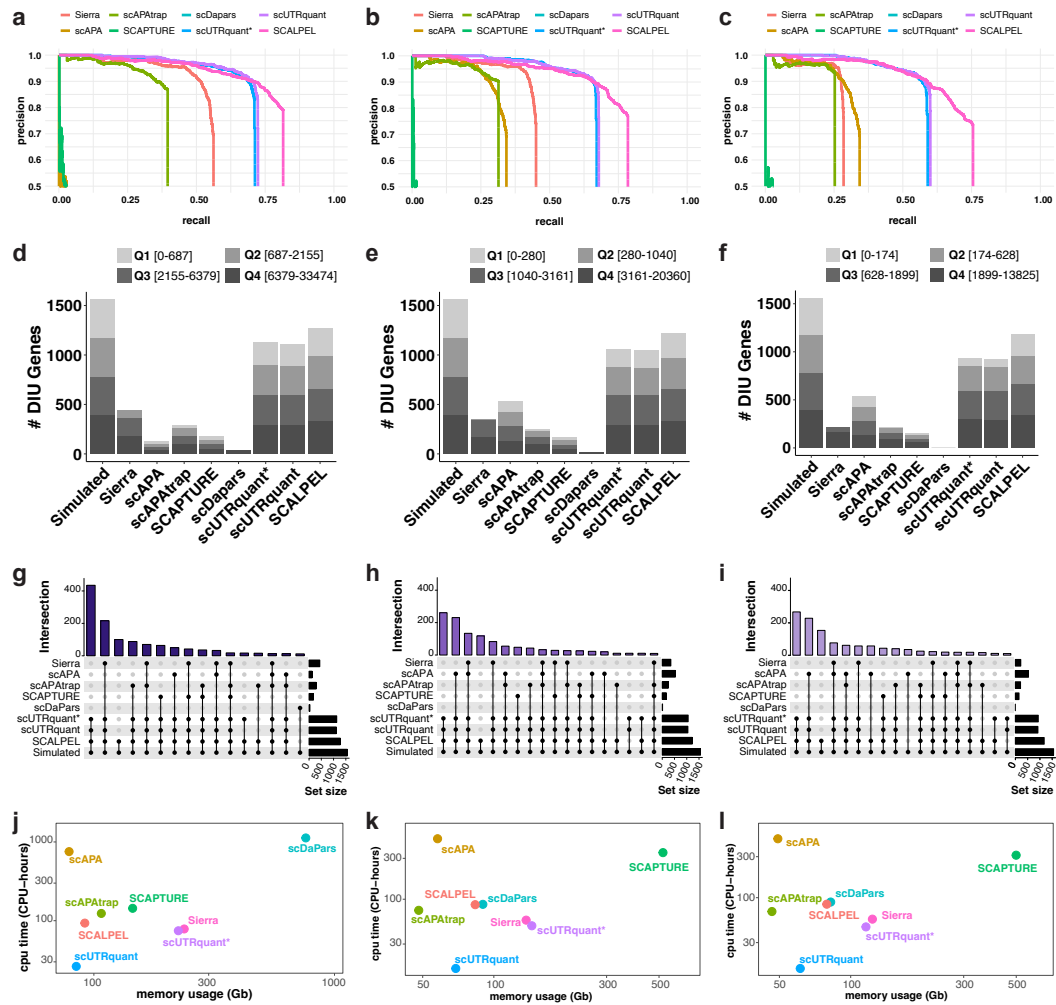
Supplementary figure 7. Additional feature plots validating the identity of the identified clusters. Feature plots of known markers that have been used to identify the cell populations obtained (Figure 3a). These markers validate the identity of NPCs (HES1), cortical fate NPCs (GLI3), dorsal fate NPCs (IRX3), astrocytes (S100B), intermediate progenitor neurons (ASCL1, EOMES), immature (STMN2) and mature (SYP) neurons, and inhibitory (GAD1) and excitatory neurons (NEUROD1).



Supplementary figure 8. Gene expression correlation across samples. Heatmap showing the pseudo-bulk gene expression correlation across samples. Inside each cell, the Pearson correlation coefficient is shown. Both ACME (A1 and A2) and vivoPHIX (V1 and V2) samples show lower correlation with all the other samples. Yet, the average correlation across all samples is quite high ($R \geq 0.8$).

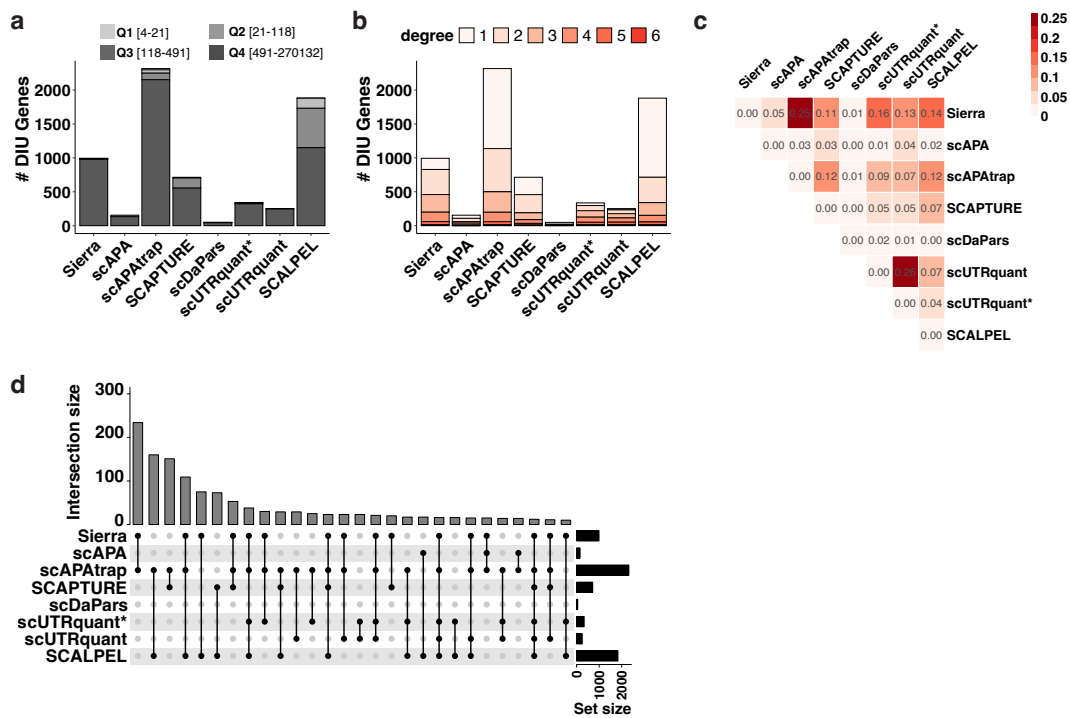


Supplementary figure 9. Cell-type expression correlation highlights fixation biases in cell clustering. Heatmap plot showing per cluster gene expression correlation across samples. Fixation method and cluster identity for each sample are indicated as color boxes on top of the heatmap. There is an overall good correlation between progenitor cell populations and excitatory neuron populations. Yet, gene expression similarity across clusters does not allow quantifying clustering biases induced by specific fixation methods.

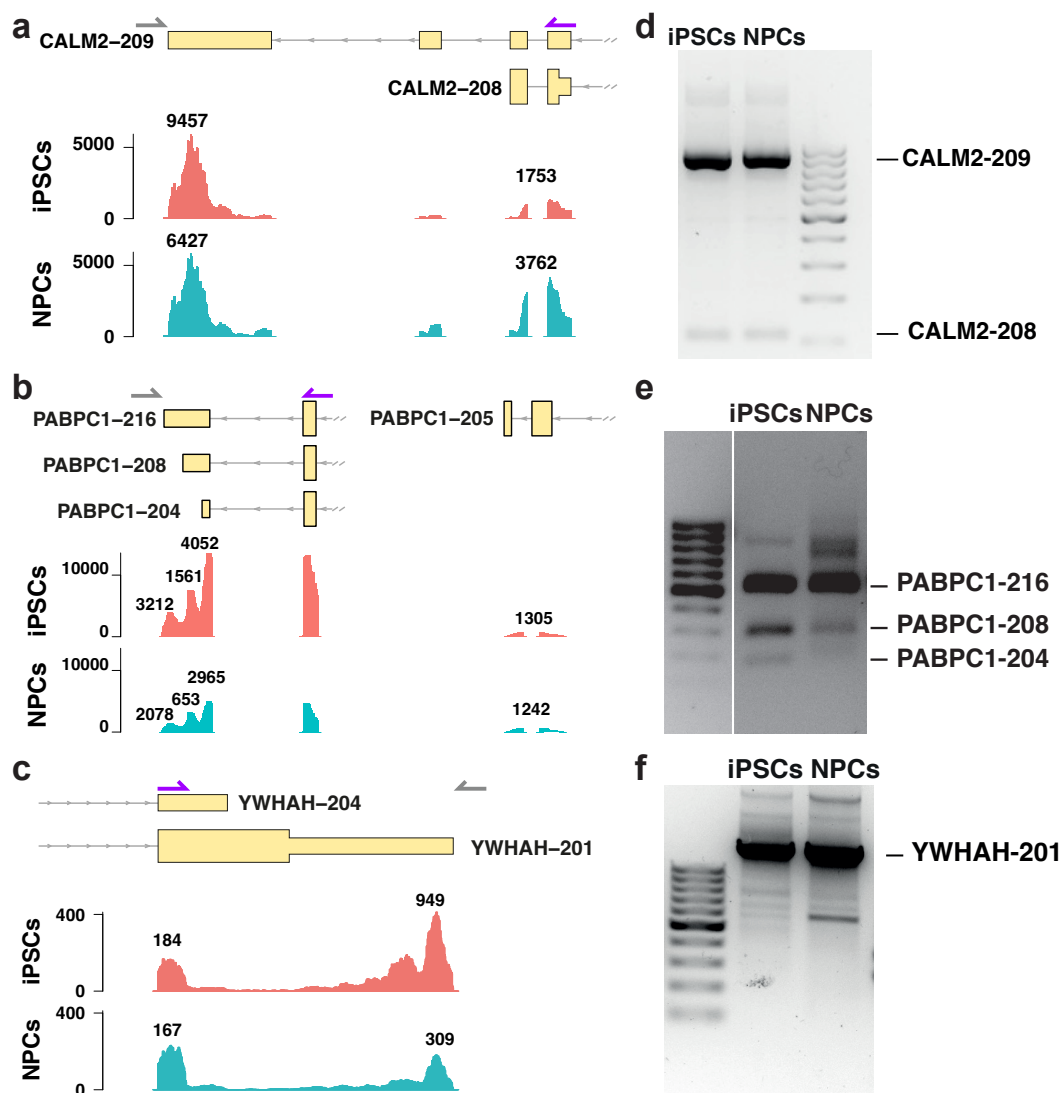


Supplementary figure 10. Additional plots on the benchmarking analysis. a-c. Precision-recall curves showing the performance of the different tools in the high (a) medium (b) and low (c) sequencing-depth datasets. In all cases SCALPEL has higher recall than all other methods while keeping a good precision. d-f. Correctly identified number of DIU genes on the high (a) medium (b) and low (c) datasets by all tools considering their expression quartile. SCALPEL correctly predicts more DIU genes than all other tools in all conditions across all quartiles. g-i. upSet plots showing the agreement in the number of correctly predicted DIU genes in the high (g) medium (h) and low (i) datasets across tools. In all cases, SCALPEL predictions show the higher degree of agreement with the predictions from other tools. j-l. Comparison of the CPU time vs the memory usage for the different tools in the high (j) medium (k) and low (l) datasets. Except for scUTRquant, which uses an already processed annotation, SCALPEL and scAPAtap are the most efficient tools (less memory and CPU time used) across datasets.

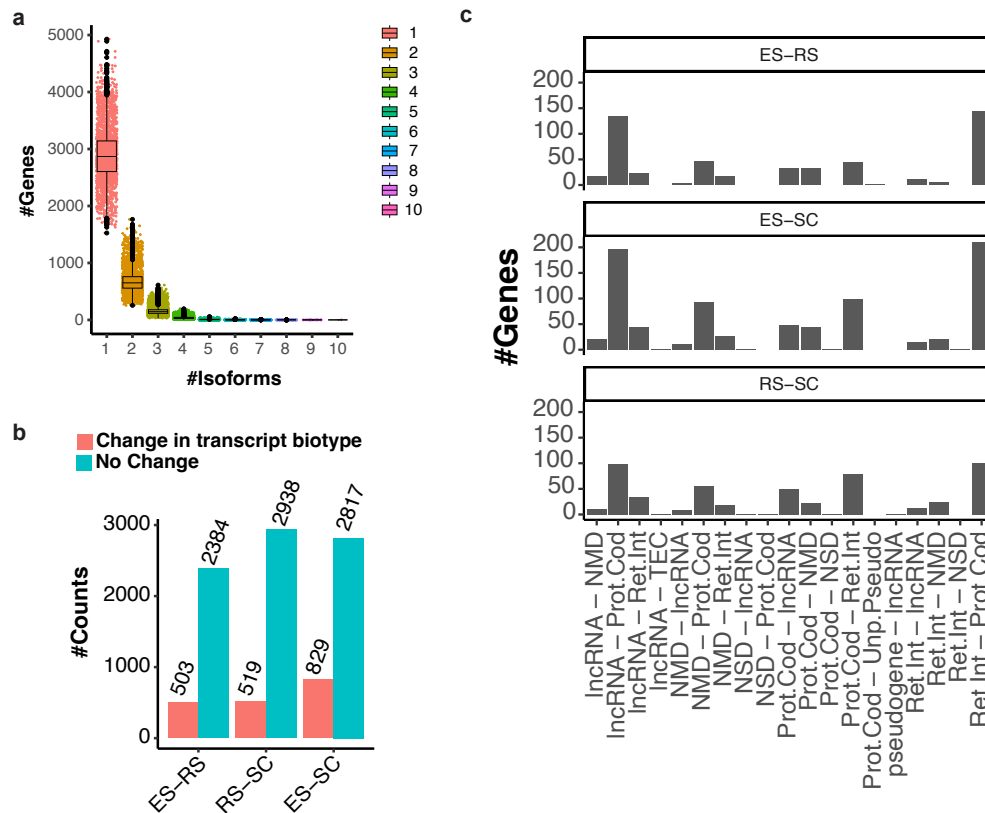
and RS (**c,f,i**), RS and SC (**d,g,j**) and SC and ES (**e,h,k**) cells. **c-d**. Distribution of predicted DIU genes across expression quartiles. In all comparisons, almost all genes belong to the quartiles with highly expressed genes (Q3 & Q4). **f-h**. Agreement in the prediction of DIU genes across tools. The agreement in the prediction of DIU genes by SCALPEL and scUTRquant using the UTRome is similar across all datasets and higher than that of other tools including scUTRquant using the standard annotation (scUTRuant*). **i-k**. UpSet plots showing the agreement in the predictions from each tool between ES and RS (**i**), RS and SC (**j**) and SC and ES (**k**). Only sets containing more than 10 genes are shown. **l-n** Heatmaps comparing the agreement in the predictions across all tool pairs in ES and RS (**l**), RS and SC (**m**), ES and SC (**n**). Scale bars show the Jaccard index coefficients for all comparisons. In all cases, SCALPEL and scUTRquant agreement with other tools is similar.



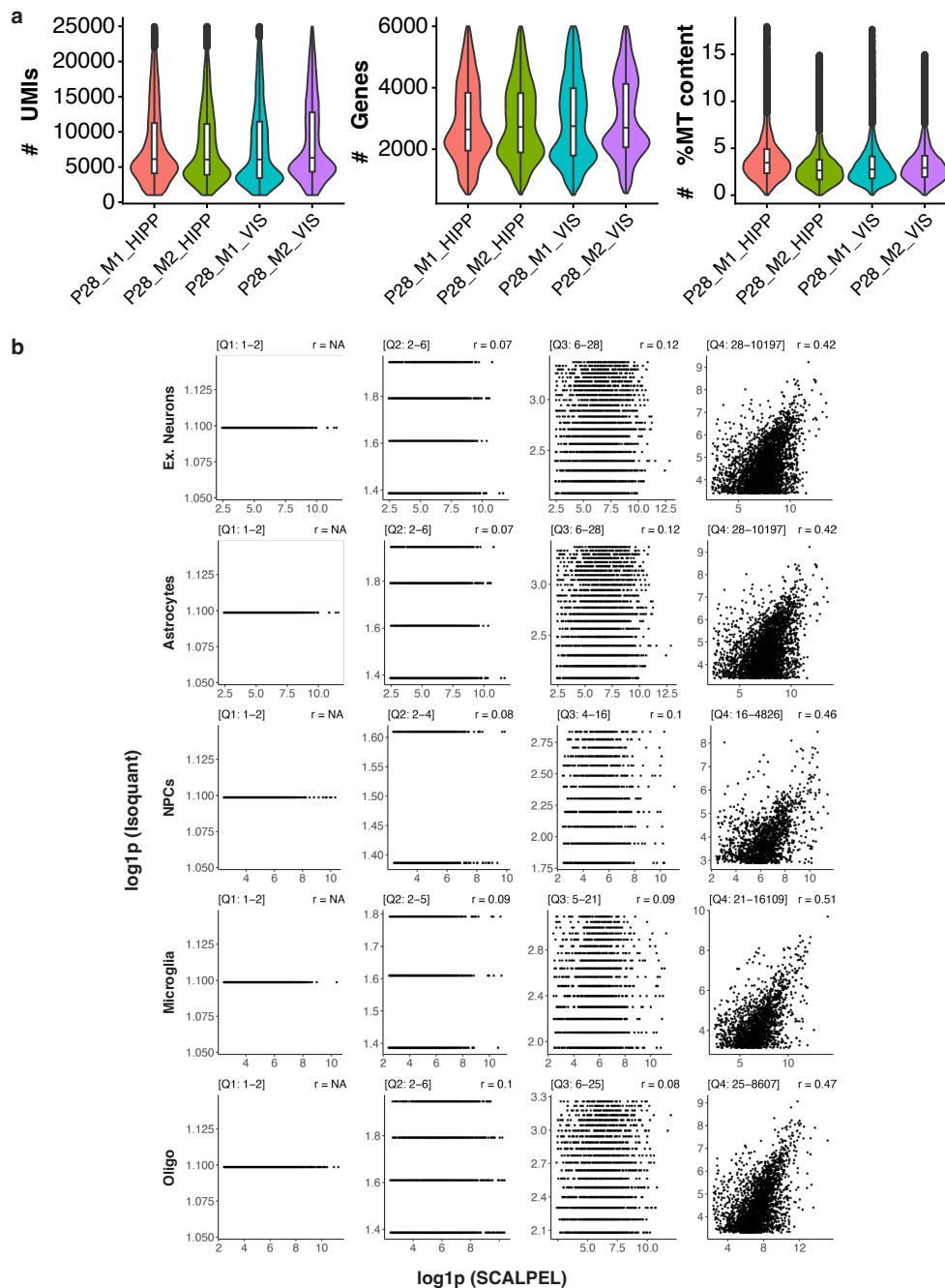
Supplementary figure 12. Benchmarking of APA and isoform quantification tools on a Drop-seq dataset containing iPSCs and NPCs. **a.** DIU genes identified by each tool between iPSC and NPCs. scAPAtrop predicts more DIU genes than all other tools, mostly among highly expressed genes. Using this dataset, the performance of scUTRquant drops dramatically in comparison to SCALPEL. **c.** Heatmap comparing the agreement in predictions across all tool pairs. The scale bar show the Jaccard index coefficients for all comparisons. In this case, Sierra predictions have a slightly higher agreement than SCALPEL predictions compared to other tools. **d.** UpSet plot showing the agreement in the predictions across all tools. Only sets containing at least 10 genes are shown.



Supplementary figure 13. Experimental validation of isoform changes between iPSCs and NPCs. **a-c.** SCALPEL isoform quantification of CALM2 (**a**), PABPC1 (**b**), and YWHAH (**c**) genes. Coverage plots show the distribution of filtered reads along isoforms. SCALPEL quantification of the different isoforms is included on top of the coverage tracks. The location of the oligo dT primer (grey) and the gene-specific primer (purple) used for experimental validations are shown on top of the gene diagrams. **d-f.** Experimental validation of detected isoform for the genes in **a-c** using 3' RACE. Using this technique we detect all the isoforms of CALM2 (**e**) and the main isoforms of PABPC1 (**f**) and YWHAH.



Supplementary figure 14. Additional information on the isoform diversity on the mouse spermatogenesis 10x dataset. **a.** Distribution of the number of isoforms detected per gene in individual cells. **b.** Barplot showing the fraction of DIU genes in which the major isoform changes the biotype and thus could lead to strong biological effects. These cases represent ~15-22% of all isoform changes. **c.** Characterization of the biotypes of the main isoform for the DIU genes in which there is a change in the biotype of the main isoform



Supplementary figure 15. Additional details on the scRNA-seq 10x datasets of mouse brain with paired long and short read data. a. Distribution of the number of UMIs, genes and mitochondrial content for each of the four samples analyzed. **b.** Correlation between the isoform quantification done by isoquant using PacBio data and the quantification done by SCALPEL using the 10x data. For each cell type, we show the correlation for the isoforms in each of the quantiles according to isoquant quantification. While highly expressed isoforms according to isoquant (Q4) have a good correlation with SCALPEL quantification, low and average expressed isoforms (Q1-Q3) have a poor correlation with SCALPEL quantification

8.2. Supplementary Tables

(Ex)

Sample	Initial number of cells	% of lost cells			
		UMI < 200 or UMI > 17000	Genes < 200 or Genes > 5500	Mitochondrial content > 7.5%	Ribosomal > 40%
F1	1764	0.68	1.54	0.17	0
F2	1327	4.75	3.48	0.25	0
M1	1841	3.48	2.98	0	0.05
M2	1752	2.45	1.93	0	0.12
M3	1833	0.22	0.49	0.11	0.11
M4	1305	0.15	3.15	0.4	0.24
D1	1234	0.65	0	0.98	0
D2	1342	1.27	0	0.83	0.3
D3	1564	1.92	0.06	1.96	0.067
A1	1313	29.09	15.79	0	1.79
A2	1524	13.58	7.9	0	2.06
V1	1573	19,83	5,15	0,08	0
V2	3053	48,74	10,54	0	0

Supplementary table 1: Cumulative percentage of cells discarded using each of the different parameters used to remove low quality cells

Sample	PCs	pK	No. of doublets
F1	11	0.02	111
F2	10	0.02	75
M1	14	0.06	110
M2	10	0.02	105
M3	11	0.04	119
M4	7	0.02	82
D1	12	0.02	75
D2	8	0.02	82
D3	10	0.25	97
A1	6	0.04	40
A2	9	0.03	68
V1	12	0.04	72
V2	11	0.06	77

Supplementary table 2: Parameters and number of doublets identified using DoubletFinder

Use Hyperlink for extended table:

https://zenodo.org/records/15889114/files/Thesis_Franz_AKE_supplementary_table.xlsx?download=1

Supplementary table 3: Differentially expressed genes per cluster for each fixation/preservation method

Type	Gene expression Changes	Differential Isoform Usage	Number of isoforms	Number of Genes	Number of Isoforms2
unique isoform genes with no significant expression changes across clusters	NO	NO	1	1000	1000
unique isoform genes with significant expression changes across clusters	YES	NO	1	1000	1000
multiple isoform genes with no significant changes in clusters	NO	NO	4	600	2400
multiple isoform genes with significant expression changes across clusters	$\log_2FC = 1.25 - 1.50 $	NO	2	800	1600
multiple isoform genes with significant expression changes across clusters	$\log_2FC = 1.50 - 2 $	NO	2	800	1600
multiple isoform genes with significant expression changes across clusters	$\log_2FC > 2 $	YES	2	800	1600
DIU genes with significant expression changes across clusters and changes in expression	$\log_2FC = 1.25 - 1.50 $	YES	2	520	1040
DIU genes with significant expression changes across clusters and changes in expression	$\log_2FC = 1.50 - 2 $	YES	2	520	1040
DIU genes with significant expression changes across clusters and changes in expression	$\log_2FC > 2 $	YES	2	520	1040
Total				6560	12320

Supplementary table 4: Types of genes simulated

Tools	PAS quantification	Quantification restricted to 3' UTR region
Sierra	Peaks	NO
scAPA	Peaks	YES
scAPATrap	Peaks	NO
scDaPars	Peaks	YES
SCAPTURE	Peaks / Isoforms	NO
scUTRquant	Peaks / Isoforms	NO
SCALPEL	Isoforms	NO

Supplementary table 5: Characteristics of tools included in the benchmarking

gene_name	cluster	avg_log2FC	pct.1	pct.2	p_val	p_val_adj
Atp1a2	Astrocytes	6,35787298	0,977	0,132	0	0
Htra1	Astrocytes	6,594727221	0,924	0,085	0	0
Slco1c1	Astrocytes	7,030668718	0,839	0,03	0	0
Prex2	Astrocytes	5,906356535	0,893	0,093	0	0
Gja1	Astrocytes	7,048844995	0,834	0,034	0	0
Gpc5	Astrocytes	6,664235834	0,94	0,15	0	0
Pla2g7	Astrocytes	6,287857726	0,827	0,051	0	0
Ntsr2	Astrocytes	6,306371947	0,815	0,044	0	0
Gpr37l1	Astrocytes	6,030823731	0,836	0,066	0	0
Bcan	Astrocytes	5,692393397	0,843	0,074	0	0
S1pr1	Astrocytes	6,2676565	0,8	0,042	0	0
F3	Astrocytes	7,038273429	0,778	0,026	0	0
Ptpz1	Astrocytes	4,657850655	0,953	0,201	0	0
Fam107a	Astrocytes	6,96661993	0,774	0,024	0	0
Plpp3	Astrocytes	5,658129093	0,957	0,218	0	0
Ndr2	Astrocytes	5,242526066	0,831	0,092	0	0
Aqp4	Astrocytes	6,99731382	0,759	0,025	0	0
Rgs20	Astrocytes	4,693625354	0,875	0,147	0	0
Aldoc	Astrocytes	5,897576215	0,851	0,124	0	0
Npas3	Astrocytes	5,05446588	0,938	0,226	0	0
Dlgap2	Excitatory.neurons	6,358130039	0,953	0,052	0	0
Lrnf5	Excitatory.neurons	5,802440792	0,945	0,059	0	0
Ryr2	Excitatory.neurons	5,788919747	0,949	0,065	0	0
Kctd16	Excitatory.neurons	5,579065054	0,948	0,076	0	0
Grin2a	Excitatory.neurons	6,096741355	0,969	0,101	0	0
Frmpd4	Excitatory.neurons	5,625891326	0,911	0,048	0	0
Sntg1	Excitatory.neurons	4,827496912	0,923	0,064	0	0
Cacna1e	Excitatory.neurons	4,809582996	0,918	0,074	0	0
Slit3	Excitatory.neurons	4,530478663	0,889	0,055	0	0
Fam135b	Excitatory.neurons	5,034471001	0,868	0,043	0	0
Epha6	Excitatory.neurons	5,758802803	0,867	0,045	0	0
Lrrc7	Excitatory.neurons	3,752936539	0,964	0,144	0	0
Kcnj6	Excitatory.neurons	5,453893255	0,849	0,031	0	0
Brinp1	Excitatory.neurons	4,88519485	0,931	0,115	0	0
Erc2	Excitatory.neurons	4,546665726	0,959	0,148	0	0
Rims1	Excitatory.neurons	4,565099418	0,954	0,146	0	0
Grm5	Excitatory.neurons	4,226430254	0,963	0,156	0	0
Unc80	Excitatory.neurons	4,551389983	0,868	0,062	0	0
Fam155a	Excitatory.neurons	3,59635626	0,957	0,152	0	0
Clstn2	Excitatory.neurons	3,351214407	0,901	0,097	0	0
Laptn5	Microglia	6,529662377	0,975	0,055	0	0

Siglech	Microglia	6,662737532	0,957	0,044	0	0
Fcer1g	Microglia	6,338394729	0,968	0,059	0	0
Fcrls	Microglia	6,523076438	0,966	0,06	0	0
Ctss	Microglia	6,604675852	0,986	0,083	0	0
Selplg	Microglia	6,421575322	0,955	0,054	0	0
Tyrobp	Microglia	6,309735016	0,976	0,077	0	0
Inpp5d	Microglia	6,626951652	0,939	0,041	0	0
Csf1r	Microglia	6,373945622	0,981	0,085	0	0
Trem2	Microglia	6,316744134	0,959	0,066	0	0
C1qc	Microglia	6,493716595	0,98	0,089	0	0
Ly86	Microglia	6,539327127	0,922	0,038	0	0
Tmem119	Microglia	6,522472097	0,925	0,045	0	0
Cx3cr1	Microglia	6,515594389	0,982	0,107	0	0
Olfml3	Microglia	6,358692653	0,903	0,042	0	0
Gpr34	Microglia	6,438203687	0,904	0,043	0	0
Vsir	Microglia	6,146151685	0,903	0,051	0	0
Rgs10	Microglia	5,952635772	0,959	0,108	0	0
Zfhx3	Microglia	5,831778868	0,963	0,112	0	0
Aif1	Microglia	6,330154694	0,891	0,041	0	0
Igfbpl1	NPCs	7,626371415	0,884	0,019	0	0
Sox11	NPCs	6,212415832	0,944	0,132	0	0
Dpysl3	NPCs	4,951215271	0,843	0,095	0	0
Nnat	NPCs	3,940583583	0,925	0,193	0	0
Tubb2b	NPCs	4,032831728	0,951	0,275	0	0
Sox4	NPCs	4,027061204	0,923	0,274	0	0
Dcx	NPCs	4,41997144	0,775	0,132	0	0
Gm17750	NPCs	4,375187734	0,708	0,073	0	0
Hmgb3	NPCs	4,376682909	0,733	0,112	0	0
Ddah2	NPCs	3,09412657	0,742	0,158	0	0
Fxyd6	NPCs	2,974686898	0,754	0,178	0	0
C530008M17Rik	NPCs	3,446385425	0,752	0,181	0	0
Cd24a	NPCs	6,715682333	0,574	0,011	0	0
Ppp1r14b	NPCs	3,106697222	0,867	0,315	0	0
Gm1673	NPCs	2,712779601	0,827	0,291	0	0
Sncaip	NPCs	2,965400604	0,669	0,133	0	0
Jpt1	NPCs	2,949752314	0,914	0,385	0	0
Serf1	NPCs	3,083533858	0,69	0,172	0	0
Ezh2	NPCs	3,471908812	0,676	0,165	0	0
Pafah1b3	NPCs	3,227552219	0,6	0,09	0	0
Prr5l	Oligodendrocytes	6,591751592	0,958	0,033	0	0
Fa2h	Oligodendrocytes	6,068789305	0,949	0,038	0	0
Gjc3	Oligodendrocytes	6,059006758	0,937	0,034	0	0

Plekhh1	Oligodendrocytes	6,24183112	0,929	0,027	0	0
Aspa	Oligodendrocytes	5,836903364	0,943	0,042	0	0
Ugt8a	Oligodendrocytes	6,078692043	0,981	0,082	0	0
Ermn	Oligodendrocytes	6,280517505	0,968	0,07	0	0
Tspan2	Oligodendrocytes	6,157764574	0,986	0,104	0	0
Bcas1	Oligodendrocytes	5,047524118	0,935	0,054	0	0
ErbB4	Oligodendrocytes	4,165179658	0,972	0,093	0	0
Lpar1	Oligodendrocytes	5,681931142	0,924	0,048	0	0
Plip	Oligodendrocytes	5,611678149	0,922	0,053	0	0
Nkx6-2	Oligodendrocytes	5,8072027	0,899	0,032	0	0
Tmem88b	Oligodendrocytes	5,853944706	0,898	0,033	0	0
Mog	Oligodendrocytes	6,034194711	0,996	0,132	0	0
D7Ert443e	Oligodendrocytes	5,410015127	0,895	0,035	0	0
Gal3st1	Oligodendrocytes	5,51599526	0,898	0,039	0	0
Phldb1	Oligodendrocytes	5,030903768	0,936	0,079	0	0
Gsn	Oligodendrocytes	5,833119121	0,983	0,126	0	0
NdrG1	Oligodendrocytes	5,847393765	0,902	0,047	0	0

Supplementary table 6: Top 20 gene DGE markers for 10x scRNA-seq data from mouse hippocampus and visual cortex

gene	transcript	cluster	avg_log2FC	pct.1	pct.2	p_val	p_val_adj
Gpr37l1	Gpr37l1-201	Astrocytes	6,37	0,798	0,059	0	0
Htra1	Htra1-201	Astrocytes	7,10	0,79	0,058	0	0
Aldoc	Aldoc-202	Astrocytes	6,02	0,813	0,11	0	0
Ntsr2	Ntsr2-201	Astrocytes	6,55	0,726	0,038	0	0
NdrG2	NdrG2-201	Astrocytes	5,72	0,757	0,074	0	0
Plpp3	Plpp3-201	Astrocytes	6,73	0,769	0,09	0	0
Bcan	Bcan-201	Astrocytes	6,19	0,733	0,054	0	0
Mt2	Mt2-201	Astrocytes	6,13	0,743	0,069	0	0
Mfge8	Mfge8-202	Astrocytes	5,82	0,722	0,068	0	0
Cldn10	Cldn10-201	Astrocytes	6,69	0,658	0,027	0	0
Clu	Clu-201	Astrocytes	5,70	0,824	0,195	0	0
Slc1a3	Slc1a3-201	Astrocytes	6,32	0,7	0,073	0	0
Pla2g7	Pla2g7-201	Astrocytes	6,35	0,654	0,032	0	0
Slc1a2	Slc1a2-202	Astrocytes	5,08	0,887	0,28	0	0
Mt2	Mt2-202	Astrocytes	5,30	0,646	0,048	0	0
Cldn10	Cldn10-202	Astrocytes	7,15	0,604	0,023	0	0
Gstm1	Gstm1-201	Astrocytes	5,39	0,63	0,072	0	0
Gja1	Gja1-201	Astrocytes	6,99	0,577	0,019	0	0

Atp1b2	Atp1b2-201	Astrocytes	4,89	0,666	0,125	0	0
Slc7a10	Slc7a10-201	Astrocytes	7,09	0,548	0,016	0	0
Kctd16	Kctd16-202	Excitatory.neurons	6,60	0,685	0,025	0	0
Slc17a7	Slc17a7-201	Excitatory.neurons	3,45	0,807	0,173	0	0
Camk2a	Camk2a-201	Excitatory.neurons	5,05	0,668	0,05	0	0
Arpp21	Arpp21-214	Excitatory.neurons	4,83	0,731	0,116	0	0
Olfm1	Olfm1-202	Excitatory.neurons	3,16	0,786	0,176	0	0
Camk2b	Camk2b-209	Excitatory.neurons	3,43	0,723	0,122	0	0
Arpp21	Arpp21-225	Excitatory.neurons	4,83	0,696	0,102	0	0
Atp1a3	Atp1a3-201	Excitatory.neurons	3,13	0,765	0,182	0	0
Slc12a5	Slc12a5-208	Excitatory.neurons	4,12	0,635	0,052	0	0
C1qtnf4	C1qtnf4-201	Excitatory.neurons	2,74	0,817	0,236	0	0
Atp1a3	Atp1a3-205	Excitatory.neurons	2,95	0,76	0,18	0	0
Snca	Snca-201	Excitatory.neurons	2,81	0,755	0,175	0	0
Camk2a	Camk2a-203	Excitatory.neurons	4,58	0,642	0,063	0	0
Olfm1	Olfm1-203	Excitatory.neurons	3,13	0,908	0,332	0	0
Ndr4	Ndr4-202	Excitatory.neurons	2,76	0,722	0,151	0	0
Atp1a3	Atp1a3-202	Excitatory.neurons	3,35	0,748	0,179	0	0
Syp	Syp-201	Excitatory.neurons	3,00	0,716	0,149	0	0
Snca	Snca-201	Excitatory.neurons	2,93	0,775	0,212	0	0
Snca	Snca-202	Excitatory.neurons	2,95	0,775	0,212	0	0
Nptxr	Nptxr-201	Excitatory.neurons	3,45	0,676	0,113	0	0
Ctss	Ctss-201	Microglia	6,60	0,976	0,063	0	0
Ctss	Ctss-202	Microglia	6,65	0,976	0,063	0	0
Fcer1g	Fcer1g-201	Microglia	6,23	0,962	0,051	0	0
Cx3cr1	Cx3cr1-202	Microglia	6,91	0,939	0,031	0	0
Tyrbp	Tyrbp-201	Microglia	6,23	0,974	0,069	0	0
Tyrbp	Tyrbp-202	Microglia	6,18	0,974	0,069	0	0
Csf1r	Csf1r-201	Microglia	6,26	0,971	0,07	0	0
C1qc	C1qc-201	Microglia	6,42	0,981	0,08	0	0
Trem2	Trem2-201	Microglia	6,29	0,952	0,058	0	0
Trem2	Trem2-202	Microglia	6,18	0,951	0,057	0	0
Cx3cr1	Cx3cr1-201	Microglia	6,32	0,948	0,061	0	0
Laptm5	Laptm5-207	Microglia	6,29	0,911	0,042	0	0
C1qa	C1qa-201	Microglia	6,44	0,985	0,123	0	0
C1qb	C1qb-201	Microglia	6,29	0,984	0,122	0	0
Tmem119	Tmem119-201	Microglia	6,40	0,9	0,039	0	0
Selpg	Selpg-205	Microglia	6,49	0,878	0,04	0	0
Siglech	Siglech-202	Microglia	6,81	0,862	0,026	0	0
Selpg	Selpg-204	Microglia	6,21	0,875	0,04	0	0
Laptm5	Laptm5-201	Microglia	6,69	0,859	0,024	0	0
Sparc	Sparc-208	Microglia	5,03	0,95	0,129	0	0

Nnat	Nnat-202	NPCs	4,12	0,889	0,106	0	0
Dpysl3	Dpysl3-202	NPCs	5,65	0,656	0,037	0	0
Dpysl3	Dpysl3-201	NPCs	5,85	0,649	0,036	0	0
Hnrnpa1	Hnrnpa1-202	NPCs	3,64	0,812	0,216	0	0
Ppp1r14b	Ppp1r14b-201	NPCs	3,20	0,852	0,279	0	0
Ddah2	Ddah2-201	NPCs	3,27	0,704	0,135	0	0
Tmsb10	Tmsb10-203	NPCs	3,53	0,942	0,375	0	0
Jpt1	Jpt1-201	NPCs	3,25	0,811	0,254	0	0
Gm1673	Gm1673-203	NPCs	3,06	0,79	0,242	0	0
Nfib	Nfib-202	NPCs	4,84	0,636	0,09	0	0
Fxyd6	Fxyd6-203	NPCs	2,95	0,696	0,151	0	0
Gm1673	Gm1673-201	NPCs	2,67	0,773	0,232	0	0
Zbtb20	Zbtb20-203	NPCs	3,24	0,748	0,209	0	0
Marcks1	Marcks1-201	NPCs	2,79	0,844	0,308	0	0
Zbtb20	Zbtb20-205	NPCs	2,89	0,733	0,206	0	0
Ddah2	Ddah2-204	NPCs	3,14	0,632	0,107	0	0
Serf1	Serf1-201	NPCs	3,34	0,661	0,15	0	0
Tmsb10	Tmsb10-204	NPCs	3,08	0,661	0,152	0	0
Hmgb3	Hmgb3-201	NPCs	4,31	0,572	0,065	0	0
Nfix	Nfix-202	NPCs	3,52	0,857	0,352	0	0
Mag	Mag-204	Oligodendrocytes	6,44	0,952	0,039	0	0
Tspan2	Tspan2-202	Oligodendrocytes	7,38	0,918	0,022	0	0
Mag	Mag-207	Oligodendrocytes	6,11	0,925	0,034	0	0
Mog	Mog-201	Oligodendrocytes	6,41	0,964	0,086	0	0
Mobp	Mobp-202	Oligodendrocytes	6,73	0,945	0,068	0	0
Mal	Mal-201	Oligodendrocytes	7,78	0,985	0,11	0	0
Gsn	Gsn-201	Oligodendrocytes	5,92	0,973	0,1	0	0
Gsn	Gsn-205	Oligodendrocytes	5,89	0,973	0,1	0	0
Tspan2	Tspan2-201	Oligodendrocytes	6,57	0,886	0,025	0	0
Mag	Mag-203	Oligodendrocytes	5,47	0,978	0,121	0	0
Ppp1r14a	Ppp1r14a-202	Oligodendrocytes	6,04	0,891	0,036	0	0
Ugt8a	Ugt8a-203	Oligodendrocytes	7,18	0,877	0,028	0	0
Apod	Apod-202	Oligodendrocytes	5,02	0,911	0,069	0	0
Olig1	Olig1-201	Oligodendrocytes	4,53	0,934	0,094	0	0
Pdlim2	Pdlim2-201	Oligodendrocytes	5,43	0,889	0,051	0	0
Pdlim2	Pdlim2-210	Oligodendrocytes	5,20	0,885	0,051	0	0
Enpp2	Enpp2-201	Oligodendrocytes	4,86	0,909	0,079	0	0
Cldn11	Cldn11-201	Oligodendrocytes	6,09	0,972	0,144	0	0
Ndr1	Ndr1-201	Oligodendrocytes	6,14	0,849	0,026	0	0
Enpp2	Enpp2-204	Oligodendrocytes	5,05	0,881	0,059	0	0

Supplementary table 7: Top20 isoform iDGE markers for 10x scRNA-seq data from mouse hippocampus and visual cortex

simulated dataset	tool	quantiles	DIU Genes	Percentage DIU Genes
high	Simulated	Q1	390	100
high	Simulated	Q2	390	100
high	Simulated	Q3	390	100
high	Simulated	Q4	390	100
high	Sierra	Q2	75	19
high	Sierra	Q3	189	48
high	Sierra	Q4	175	45
high	scAPA	Q1	22	6
high	scAPA	Q2	37	9
high	scAPA	Q3	32	8
high	scAPA	Q4	34	9
high	scAPAttrap	Q1	29	7
high	scAPAttrap	Q2	81	21
high	scAPAttrap	Q3	84	22
high	scAPAttrap	Q4	96	25
high	SCAPTURE	Q1	41	11
high	SCAPTURE	Q2	43	11
high	SCAPTURE	Q3	46	12
high	SCAPTURE	Q4	48	12
high	scDaPars	Q3	1	0
high	scDaPars	Q4	35	9
high	scUTRquant*	Q1	233	60
high	scUTRquant*	Q2	301	77
high	scUTRquant*	Q3	302	77
high	scUTRquant*	Q4	293	75
high	scUTRquant	Q1	226	58
high	scUTRquant	Q2	296	76
high	scUTRquant	Q3	299	77
high	scUTRquant	Q4	291	75
high	SCALPEL	Q1	284	73
high	SCALPEL	Q2	330	85
high	SCALPEL	Q3	330	85
high	SCALPEL	Q4	328	84
mid	Simulated	Q1	390	100
mid	Simulated	Q2	390	100
mid	Simulated	Q3	390	100
mid	Simulated	Q4	390	100

mid	Sierra	Q2	7	2
mid	Sierra	Q3	177	45
mid	Sierra	Q4	168	43
mid	scAPA	Q1	111	28
mid	scAPA	Q2	146	37
mid	scAPA	Q3	148	38
mid	scAPA	Q4	130	33
mid	scAPATrap	Q1	16	4
mid	scAPATrap	Q2	58	15
mid	scAPATrap	Q3	74	19
mid	scAPATrap	Q4	98	25
mid	SCAPTURE	Q1	35	9
mid	SCAPTURE	Q2	42	11
mid	SCAPTURE	Q3	39	10
mid	SCAPTURE	Q4	50	13
mid	scDaPars	Q4	14	4
mid	scUTRquant*	Q1	186	48
mid	scUTRquant*	Q2	278	71
mid	scUTRquant*	Q3	302	77
mid	scUTRquant*	Q4	293	75
mid	scUTRquant	Q1	179	46
mid	scUTRquant	Q2	277	71
mid	scUTRquant	Q3	300	77
mid	scUTRquant	Q4	291	75
mid	SCALPEL	Q1	257	66
mid	SCALPEL	Q2	309	79
mid	SCALPEL	Q3	333	85
mid	SCALPEL	Q4	325	83
low	Simulated	Q1	390	100
low	Simulated	Q2	390	100
low	Simulated	Q3	390	100
low	Simulated	Q4	390	100
low	Sierra	Q3	61	16
low	Sierra	Q4	161	41
low	scAPA	Q1	111	28
low	scAPA	Q2	146	37
low	scAPA	Q3	148	38
low	scAPA	Q4	130	33
low	scAPATrap	Q1	9	2
low	scAPATrap	Q2	51	13
low	scAPATrap	Q3	64	16
low	scAPATrap	Q4	88	23

low	SCAPTURE	Q1	20	5
low	SCAPTURE	Q2	34	9
low	SCAPTURE	Q3	41	11
low	SCAPTURE	Q4	55	14
low	scUTRquant*	Q1	86	22
low	scUTRquant*	Q2	262	67
low	scUTRquant*	Q3	295	76
low	scUTRquant*	Q4	293	75
low	scUTRquant	Q1	76	19
low	scUTRquant	Q2	258	66
low	scUTRquant	Q3	296	76
low	scUTRquant	Q4	291	75
low	SCALPEL	Q1	221	57
low	SCALPEL	Q2	299	77
low	SCALPEL	Q3	325	83
low	SCALPEL	Q4	333	85

Supplementary table 8: DIU genes identified by each benchmarked tool

Tools	Dataset	Execution time (CPU-hours)	CPU defined	Memory usage
Sierra	High	78,5	100	237,4
scAPA	High	754	100	78,9
scAPATrap	High	123,6	100	107,4
scDaPars	High	143,58	100	144925
SCAPTURE	High	1124,86	100	759,8
scUTRquant*	High	74,5	100	224,4
scUTRquant	High	26,28	100	84,5
SCALPEL	High	93,19	100	91,7
Sierra	Mid	57	100	137
scAPA	Mid	508	100	57,8
scAPATrap	Mid	74,2	100	48,1
scDaPars	Mid	87,16	100	89,85
SCAPTURE	Mid	349,72	100	519,6
scUTRquant*	Mid	49	100	144,8
scUTRquant	Mid	15,56	100	69
SCALPEL	Mid	86,38	100	83,4
Sierra	Low	56,25	100	123,5
scAPA	Low	485	100	49,25
scAPATrap	Low	69	100	46,5
scDaPars	Low	88,95	100	82,31
SCAPTURE	Low	312,14	100	499,5
scUTRquant*	Low	45,7	100	116
scUTRquant	Low	15	100	61,25
SCALPEL	Low	83,97	100	79,25

Supplementary table 9: Execution performance on the simulated scRNA-seq dataset

Tools	Total PAS / isoforms	Total Genes	multiple PAS / Isoforms Genes	DIU genes in ES-RS	DIU genes in RS-SC	DIU genes in ES-SC	Execution time (CPU hours)	CPU defined	Memory usage
Sierra	15.629	8.854	3.802	1.555	1.773	2.036	56.2	40	139 Gb
scAPA	9.775	7.678	1.711	323	655	532	636	40	19 Gb
scAPATrap	41.998	14.093	8.824	1.756	2.385	2.867	68.8	40	39.16 Gb
SCAPTURE	15.337	13.168	2.039	716	970	1.178	262.7	40	707 Gb
scDaPars	3.300	2.241	2.241	80	102	223	137.4	40	48 Gb
scUTRquant*	34.188	16.558	9.174	1.468	1.678	1.831	22.4	40	118 Gb
scUTRquant	36.016	16.475	10.225	2.434	2.674	2.736	12.3	40	22 Gb
SCALPEL	51.767	17.525	10.577	2.889	3.458	3.649	49.60	40	42 Gb

Supplementary table 10: Details on the number of predictions and technical details on the mouse spermatogenesis 10x dataset

Use Hyperlink for extended table:

https://zenodo.org/records/15889114/files/Thesis_Franz_AKE_supplementary_table.xlsx?download=1

Supplementary table 11: DIU genes between iPSCs and NPCs

Use Hyperlink for extended table:

https://zenodo.org/records/15889114/files/Thesis_Franz_AKE_supplementary_table.xlsx?download=1

Supplementary table 12: Gene markers on the high-resolution clustering analysis of mouse spermatogenesis dataset

Use Hyperlink for extended table:

https://zenodo.org/records/15889114/files/Thesis_Franz_AKE_supplementary_table.xlsx?download=1

Supplementary table 13: Isoform markers on the high-resolution clustering analysis of mouse spermatogenesis dataset

Term	Adjusted. P.value	Combine d.Score	Genes
Cilium Organization (GO:0044782)	1,13E-07	75,3750459	ARF4;CEP126;TRAF3IP1;PKD2;PTPDC1;RPGR;CATIP;TBC1D31;WDR90;TCTN2;TCTN1;TMEM107;TEKT3;TEKT4;BBS7;CFAP43;CEP290;DYNC2H1;CCDC113;RFX3;IFT80;ARMC2;IFT88;IFT43;RILPL2;IFT81;NME8;TMEM216;CFAP58;CCNO;CFAP1;ARL6;ASAP1;SLC9A3R1;ARL13B;STK36;POC1B;TOGARAM1;MAP4;PIBF1;BBS2;WDR19;SCLT1;FOXJ1;CPLANE1;WDR54;ATAT1;TMEM17;TTLL3;CEP83;CCDC5;CEP41;CCDC66;DZIP1;CLUAP1;TTLL8
Cilium Assembly (GO:0060271)	7,74E-07	61,768754	ARF4;CEP126;LRGUK;TRAF3IP1;PTPDC1;RPGR;TBC1D31;WDR90;TCTN2;TCTN1;TMEM107;TEKT3;TEKT4;BBS7;CFAP43;CEP290;DYNC2H1;CCDC113;RFX3;CSNK1D;IFT80;IFT88;IFT43;RILPL2;IFT81;NME8;TMEM216;CFAP58;CCNO;ARL6;ASAP1;CFAP206;SPATA6;DYNC2LI1;ARL13B;STK36;ARL13A;POC1B;TOGARAM1;PIBF1;BBS2;WDR19;SCLT1;FOXJ1;CPLANE1;WDR54;ATAT1;TMEM17;TTLL3;CEP83;CCDC65;CEP41;CCDC66;DZIP1;CLUAP1;TTLL8
Plasma Membrane Bound Cell Projection Assembly (GO:0120031)	5,68E-05	41,155341	ARF4;CEP126;TRAF3IP1;PTPDC1;RPGR;TBC1D31;WDR90;TCTN2;TCTN1;TMEM107;TEKT3;TEKT4;BBS7;CFAP43;CEP290;DYNC2H1;CCDC113;RFX3;IFT80;IFT88;IFT43;RILPL2;IFT81;NME8;TMEM216;FSCN1;CFAP58;CCNO;VCL;ARL6;ASAP1;SLC9A3R1;RAP1A;ARL13B;STK36;POC1B;TOGARAM1;PIBF1;BBS2;CARMIL1;WDR19;SCLT1;FOXJ1;CPLANE1;WDR54;ATAT1;GOLPH3;TMEM17;TTLL3;CEP83;CCDC65;CEP41;CCDC66;DZIP1;CLUAP1;TTLL8
Organelle Assembly (GO:0070925)	0,00804456	23,2764658	ARF4;CEP126;TRAF3IP1;PTPDC1;RPGR;TBC1D31;WDR90;TCTN2;TCTN1;TMEM107;TEKT3;TEKT4;BBS7;CFAP43;CEP290;DYNC2H1;GABARAPL2;CCDC113;RFX3;IFT80;WIPI1;ATG13;IFT88;IFT43;RILPL2;IFT81;NME8;TMEM216;CFAP58;ULK2;CCNO;ARL6;ASAP1;ARL13B;STK36;UBQLN1;POC1B;TOGARAM1;PIBF1;BBS2;WDR19;AP3D1;SCLT1;FOXJ1;CPLANE1;WDR54;ATAT1;TMEM17;TTLL3;CEP83;CCDC65;CEP41;CCDC66;DZIP1;CLUAP1;TTLL8

Supplementary table 14: GO-term enrichment analysis in RS6 vs other RS cells

Use Hyperlink for extended table:

https://zenodo.org/records/15889114/files/Thesis_Franz_AKE_supplementary_table.xlsx?download=1

Supplementary table 15: DIU genes between RS6 population and other RS populations

Use Hyperlink for extended table:

https://zenodo.org/records/15889114/files/Thesis_Franz_AKE_supplementary_table.xlsx?download=1

Supplementary table 16: DIU genes identified in the mouse spermatogenesis dataset

miRNA	Seed sequence	Number of target genes	Number of miRNA targeted isoforms	Number of non-targeted isoforms	p-value	adjusted p-value
hsa-miR-128-5p	GGGGCC	709	1105	1356	7.18E-07	0.0000115
hsa-let-7b-5p	GAGGUA	431	657	880	0.000013	0.00010376
hsa-miR-153-5p	CAUUUU	35	57	76	0.000386	0.002058667
hsa-miR-34a-5p	GGCAGU	11	18	18	0.0009875	0.00395
hsa-miR-199a-3p	CAGUAG	132	204	280	0.00247	0.007904
hsa-miR-124-3p	AAGGCA	314	452	687	0.003425	0.009133333
hsa-miR-128-3p	CACAGU	302	455	633	0.008633	0.017266
hsa-miR-9-5p	CUUUGG	436	660	932	0.007851	0.017266
hsa-miR-137-5p	CGGGUA	24	42	41	0.09643	0.171431111
hsa-miR-199a-5p	CCAGUG	54	87	111	0.1088	0.17408
hsa-miR-34a-3p	AAUCAG	76	111	162	0.1324	0.192581818
hsa-let-7b-3p	UAUACA	62	94	141	0.1452	0.1936
hsa-miR-184	GGACGG	151	217	312	0.4506	0.554584615
hsa-miR-124-5p	GUGUUC	31	49	62	0.9194	0.9365
hsa-miR-137-3p	UAUUGC	6	9	12	0.8506	0.9365
hsa-miR-153-3p	UGCAUA	38	51	88	0.9365	0.9365

Supplementary table 17: NPC miRNAs evaluated as isoform expression regulators in NPCs

Clusters	Total Genes expressed (Isoquant)	Total Isoforms expressed (Isoquant)	Total Genes expressed (SCALPEL)	Total Isoforms expressed (SCALPEL)	Number of Isoform exclusively detected by Isoquant	Number of Isoform exclusively detected by SCALPEL
Astrocytes	8.238	12.653	11.554	21.060	5.826	14.233
Microglia	8.926	14.532	11.922	22.623	6.419	14.510
Oligodendrocytes	12.429	24.693	13.379	29.197	11.470	15.974
NPCs	9.836	17.819	12.258	25.303	7.832	15.316
Excitatory neurons	7.971	12.141	11.278	20.492	5.484	13.835

Comparison	Total DIU genes	Total DIU isoforms
Excitatory neurons - Microglia	1598	3527
Excitatory neurons - Oligodendrocytes	1908	4192
Excitatory neurons - Astrocytes	1257	2773
Excitatory neurons - NPCs	1338	2978
Microglia - Oligodendrocytes	1205	2645
Microglia - Astrocytes	1071	2364
Microglia - NPCs	999	2239
Astrocytes - NPCs	811	1821
Oligodendrocytes - Astrocytes	1188	2628
Oligodendrocytes - NPCs	973	2167

Supplementary table 18: Stats of detected feature from 10x short read and PacBio long read data

Use Hyperlink for extended table:

https://zenodo.org/records/15889114/files/Thesis_Franz_AKE_supplementary_table.xlsx?download=1

Supplementary table 19: Top 20 gene DGE markers for 10x scRNA-seq data from mouse hippocampus and visual cortex

Use Hyperlink for extended table:

https://zenodo.org/records/15889114/files/Thesis_Franz_AKE_supplementary_table.xlsx?download=1

Supplementary table 20: Top20 isoform iDGE markers for 10x scRNA-seq data from mouse hippocampus and visual cortex

Use Hyperlink for extended table:

https://zenodo.org/records/15889114/files/Thesis_Franz_AKE_supplementary_table.xlsx?download=1

Supplementary table 21: Differential isoform usage (DIU) genes between the different cell-types using SCALPEL in 10x scRNA-seq data from Joglekar et al.

8.3. Publications

Article 1:

Gutiérrez-Franco, A., Ake, F., Hassan, M.N. *et al.* Methanol fixation is the method of choice for droplet-based single-cell transcriptomics of neural cells. *Commun Biol* **6**, 522 (2023). <https://doi.org/10.1038/s42003-023-04834-x>

Article 2:

Ake, F., Schilling, M., Fernández-Moya, S.M. *et al.* Quantification of transcript isoforms at the single-cell level using SCALPEL. *Nat Commun* **16**, 6402 (2025). <https://doi.org/10.1038/s41467-025-61118-0>