Quantification of alternative 3′UTR isoforms from single cell RNA-seq data with scUTRquant

Mervin M. Fansler[1,2], Gang Zhen[2], and Christine Mayr[1,2*]

[1]Tri-Institutional Training Program in Computational Biology and Medicine, Weill-Cornell Graduate College, New York, NY 10021, USA

[2]Cancer Biology and Genetics Program, Memorial Sloan Kettering Cancer Center, New York, NY

[*]Correspondence: mayrc@mskcc.org

Abstract

Although half of human genes use alternative polyadenylation (APA) to generate mRNA isoforms that encode the same protein but differ in their 3′UTRs, most single cell RNA-sequencing (scRNA-seq) pipelines only measure gene expression. Here, we describe an open-access pipeline, called scUTRquant (https://github.com/Mayrlab/scUTRquant), that measures gene and 3′UTR isoform expression from scRNA-seq data obtained from known cell types in any species. scUTRquant-derived gene and 3′UTR transcript counts were validated against standard methods which demonstrated their accuracy. 3′UTR isoform quantification was substantially more reproducible than previous methods. scUTRquant provides an atlas of high-confidence 3′ end cleavage sites at single-nucleotide resolution to allow APA comparison across mouse datasets. Analysis of 120 mouse cell types revealed that during differentiation genes either change their expression or they change their 3′UTR isoform usage. Therefore, we identified thousands of genes with 3′UTR isoform changes that have previously not been implicated in specific biological processes.

Introduction

Most transcriptome analyses performed to date quantify gene expression. However, approximately half of human genes use alternative cleavage and polyadenylation (APA) to generate mRNA isoforms that encode the same protein but that differ in their 3′ untranslated regions (3′UTRs) [1]. Moreover, many genes use intronic polyadenylation (IPA) signals to generate mRNA isoforms with alternative last exons, thus producing different protein isoforms [2-5]. APA is developmentally regulated and is dysregulated in disease [6,7]. Alternative 3′UTRs are rich in regulatory elements, including binding sites for microRNAs and RNA-binding proteins and regulate processes at the mRNA level, including localization, stability, and translation [8-15]. They also impact processes that occur co-translationally as they facilitate 3′UTR-dependent protein complex assembly, thus regulating protein localization and protein function [15-20].

The current gold standard for quantifying alternative 3′UTR isoform expression are bulk 3′ end sequencing methods [1,12,21-28]. However, many different library preparation protocols and computational pipelines exist that impede cross-dataset comparisons. As these methods require substantial amounts of material, they have largely been limited to the analysis of cell lines, complex tissues, or primary immune cell populations [1-3,12,21-29]. However, there is a great need

for APA analysis tools that use publicly available datasets as several programs were developed that estimate APA from bulk RNA-seq data [30-32]. However, these existing methods have substantial limitations, e.g. they are prone to artifacts due to uneven read coverage in 3′UTRs [33].

Single cell RNA sequencing (scRNA-seq) has revolutionized gene expression analysis as it allows to dissect gene expression profiles of individual cells. In addition to protocols that generate full-length mRNA read coverage, most datasets are generated using droplet-based methods that incorporate a unique molecular identifier (UMI) at the mRNA 3′ end for transcript counting [34]. As these protocols are conceptually similar to bulk 3′ end sequencing methods, several analysis protocols have recently been developed to quantify alternative 3′UTR isoform expression from scRNA-seq datasets in known cell types [35-44]. However, none of the pipelines were validated transcriptome-wide with respect to accuracy of the detected 3′UTR isoform expression. Moreover, nearly all use *de novo* peak calling from their dataset of interest which is prone to internal priming artifacts. 10x Genomics reads do not span cleavage sites, most pipelines cannot directly identify the corresponding mRNA 3′ ends.

We present scUTRquant, a reusable open-access Snakemake pipeline that provides single-nucleotide resolution on 3′ end cleavage sites obtained from scRNA-seq data. scUTRquant simultaneously measures gene expression and alternative 3′UTR isoform expression from given cell types or cell states. Gene counts obtained by scUTRquant correlate very well with gene counts obtained by CellRanger. scUTRquant-derived 3′UTR isoform counts also correlate strongly with data obtained from bulk 3′ end sequencing methods, demonstrating that scUTRquant is accurate. We also demonstrate that scUTRquant-derived 3′UTR isoform counts obtained from biological replicates are substantially more reproducible than when obtained from previous methods. Simultaneous assessment of gene and alternative 3′UTR isoform usage between cell types revealed that gene and 3′UTR isoform usage are independent parameters of gene regulation. Correct quantification of APA at high resolution will allow to study the regulation and function of alternative 3′UTR isoforms in any biological context or species.

Results

Atlas of functional mRNA 3′ end cleavage sites of the mouse transcriptome

Our goal was to use 3′-tagged scRNA-seq datasets to identify cell type- and condition-specific changes in 3′UTR isoform expression. To compare alternative 3′UTR isoform expression in individual cell types from different datasets, we set out to generate an atlas of high-confidence mRNA 3′ end cleavage sites of the mouse transcriptome. Bulk 3′ end sequencing methods unambiguously identify mRNA 3′ ends because the majority of reads traverse the cleavage sites and contain adenosines not present in the genome, which are indicative of poly(A) tails [1,12,21-28]. Although, 3′-tagged scRNA-seq protocols are conceptually similar, most of the reads generated by the 10x Genomics platform map upstream of 3′ end cleavage sites and only a small minority of reads contain untemplated adenosines (Fig. 1a). This feature makes it difficult to obtain single-nucleotide resolution for mRNA 3′ end cleavage sites directly from scRNA-seq data. It is therefore currently necessary to develop programs that model read distribution and assign upstream reads to the correct cleavage sites [38].

We observed that Microwell-seq generates longer reads with 40% of them traversing mRNA 3′ end cleavage sites (Fig. 1a and Supplementary Fig. 1a) [45]. As Microwell-seq was applied to 400,000 single cells derived from all major mouse organs, it allowed us to comprehensively identify mRNA 3′ end cleavage sites at single-nucleotide resolution (Fig. 1a). To generate a reference atlas of cleavage sites, we required secondary evidence for the cleavage sites to be retained. All cleavage sites that overlapped mRNA 3′ ends present in GENCODE version M21

or that were present in PolyASite (score >=3), which is the most comprehensive database of mRNA 3′ ends to date, were kept [29]. The remaining cleavage sites were filtered using the Bioconductor package cleanUpdTSeq to identify and exclude cleavage sites likely derived from mis-priming at internal poly(A) stretches [46]. Next, we added the additional cleavage sites to the GENCODE annotation if they mapped within 5,000 nucleotides downstream of known mRNA 3′ ends. This strategy added 9,487 cleavage sites belonging to 5,931 protein-coding genes to the GENCODE annotation. Overall, our universe of mRNA 3′ ends from mouse contains 46,501 cleavage sites that are present in 21,791 protein-coding genes (Supplementary Table 1). It is optional to use our reference atlas of mRNA 3′ end cleavage sites, as in the scUTRquant pipeline any transcriptome annotation can be used (Fig. 1b).

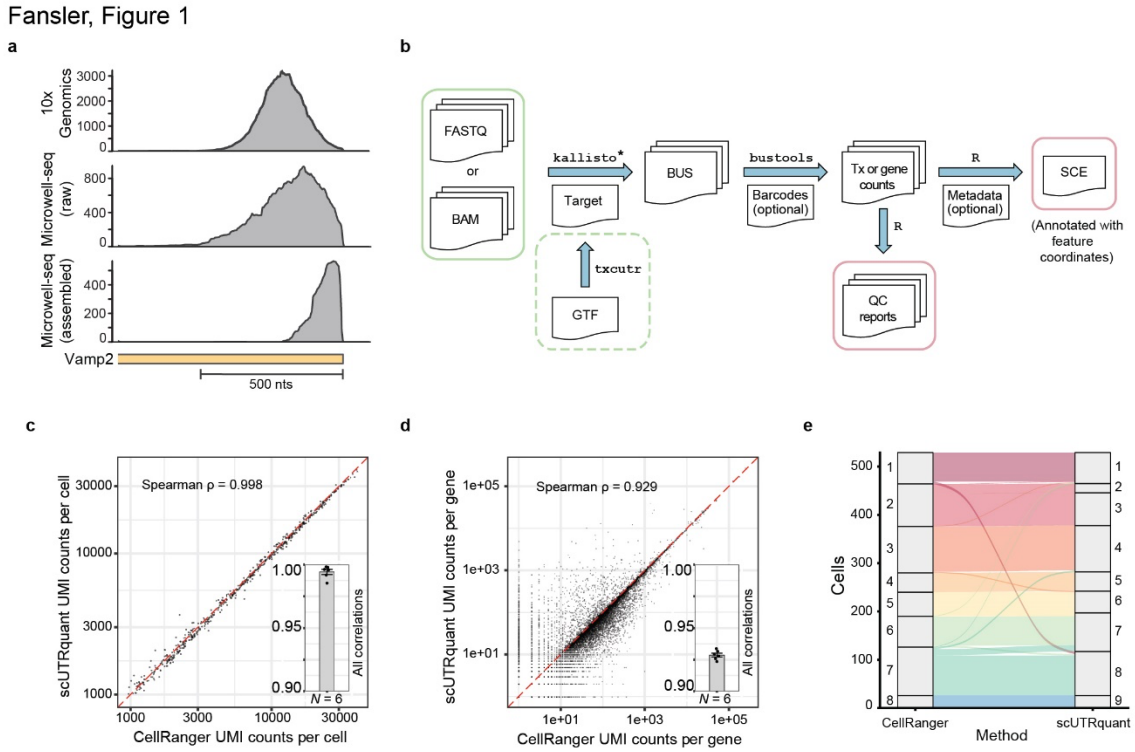

Figure 1. scUTRquant accurately quantifies gene expression.

a, Read coverage distribution at mRNA 3′ ends obtained by scRNA-seq methods shown for a representative gene. Top, 10x Genomics reads; middle, raw Microwell-seq reads; bottom, Microwell-seq reads with overlap of read1 and read2. Shown is the terminal exon of Vamp2. Nts, nucleotides.

b, Schematic of the scUTRquant Snakemake pipeline. The pipeline takes as input (green box) either FASTQ or BAM formatted files. A customized build of kallisto is used to pseudoalign against a default target index for mouse. Alternatively, the txcutr package can be used to generate a customized truncated target transcriptome, including starting from an arbitrary GTF file. An intermediate BUS file is generated and then processed by bustools, which can optionally accept a barcode whitelist. This outputs gene and isoform counts, which are processed by R to generate quality control (QC) reports and a SingleCellExperiment (SCE) object with optionally attached metadata, such as known cluster assignments or sample metadata.

c, Correlation of UMIs per cell obtained by scUTRquant and CellRanger for six mouse 10x Genomics demonstration datasets. The Spearman's ρ is shown.

d, Same as **c**, but UMIs per gene are plotted.

e, Alluvial diagram comparing Louvain clustering results derived from gene counts obtained by CellRanger and scUTRquant for the 10x Genomics demonstration dataset mouse Heart_1k_v3 (ARI = 0.84).

Transcript quantification with a truncated UTRome

To quantify alternative 3′UTR isoforms from scRNA-seq data, we built upon the kallisto-bustools toolset [47,48]. The kallisto tool pseudoaligns reads to a reference transcriptome and allows us to align the data to a given transcriptome rather than to the genome which would require a splice-aware peak caller for quantification [37]. Quantification of alternative transcript isoforms by kallisto works best if the transcripts contain unique regions [47]. However, within terminal exons, the sequences of short 3′UTR (SU) isoforms are fully contained within long 3′UTR (LU) isoforms, thus making these regions ambiguous for quantification. To minimize the overlap between alternative 3′UTR transcripts, we generated a truncated UTRome that contains 500 nucleotides of sequence upstream of all functional mRNA 3′ end cleavage sites.

The cut-off for the truncation was empirically determined (Supplementary Fig. 1b-d). We observed that more than 99% of UMIs of tested reference genes map within 500 nucleotides upstream of cleavage sites (Supplementary Fig. 1b-d). Next, we performed simulations to identify the minimum distance between cleavage sites that allows accurate 3′UTR isoform quantification if the distance between alternative cleavage sites is smaller than 500 nucleotides. This revealed that counts from isoforms that fall within 200 nucleotides may not be quantified accurately (Supplementary Fig. 1e). Therefore, when performing 3′UTR isoform quantification, we have kallisto merge isoforms that fall within 200 nucleotides and report the merged result labeled as the most distal cleavage site. This strategy allows us to quantify all 3′UTR isoforms that are further apart.

Quantification of gene expression using scUTRquant is accurate

The scUTRquant pipeline simultaneously quantifies gene expression as well as 3′UTR isoform expression (Fig. 1b). To test if scUTRquant measures gene expression accurately, we compared it to currently used standard methods, such as CellRanger [49]. We used six 10x Genomics demonstration datasets and correlated the UMI counts per cell obtained by CellRanger and by scUTRquant. We obtained near perfect correlations with Spearman's rank correlation coefficients (ρ) of greater than 0.99 (Fig. 1c). When comparing UMI counts per gene obtained by the two methods, we again observed strong correlations with Spearman's ρ of 0.93, indicating that scUTRquant is accurate with respect to gene counts (Fig. 1d). Most downstream scRNA-seq analyses use clustering to identify groups of cells with similar gene expression patterns. Therefore, we compared the Louvain clustering results based on gene counts obtained by CellRanger and scUTRquant (Fig. 1e). This revealed highly similar clustering results with adjusted RAND index (ARI) values of up to 0.89 (Table 1), thus indicating that scUTRquant and CellRanger can be used interchangeably to measure gene expression from scRNA-seq data.

Quantification of 3′UTR isoform expression using scUTRquant is accurate

The current gold standard for quantifying 3′UTR isoform expression are bulk 3′ end sequencing methods [1,12,21-28]. To assess if scUTRquant-derived 3′UTR isoform counts reflect the actual isoform expression in cells, we compared them to bulk 3′ end sequencing counts obtained from the same cell type. As 3′ end sequencing methods require a lot of material, most available datasets contain complex tissues or cell lines [29]. Few datasets have quantified 3′UTR isoform expression in individual well-defined primary cell types, such as embryonic stem cells (ESC) or hematopoietic stem cells (HSC), and were amenable for comparison [26-28,50-53].

When comparing 3′UTR transcript counts obtained from bulk 3′ end sequencing methods with scUTRquant for FACS-sorted HSCs, we observed a strong correlation (Fig. 2a, Spearman's ρ= 0.88) [28,50,51]. For ESC, the correlation was less strong (Fig. 2b, Spearman's ρ= 0.72). This was likely due to the fact that different conditions were used to keep ESC undifferentiated [26,27,52,53].

3′UTR isoforms can either be characterized by expression values of the individual isoforms or as a single value for each gene such as the long 3′UTR index (LUI) which reflects 3′UTR isoform ratio or isoform usage. LUI is the fraction of reads that map to the LU isoform out of all reads that map to the 3′UTR. LUI values for a gene range from 0 to 1 and high LUI values characterize genes with predominant LU isoform expression. When LUI values were used for comparison between 3′ end sequencing methods and scUTRquant, slightly lower correlations were observed with Spearman's ρ of up to 0.79 (Supplementary Table 2). Still, we consider the level of correlation between the current gold standard method and scUTRquant transcript counts as excellent, considering that the procedures were performed by different laboratories using vastly different methods [26-28,50-53].

Fansler, Figure 2

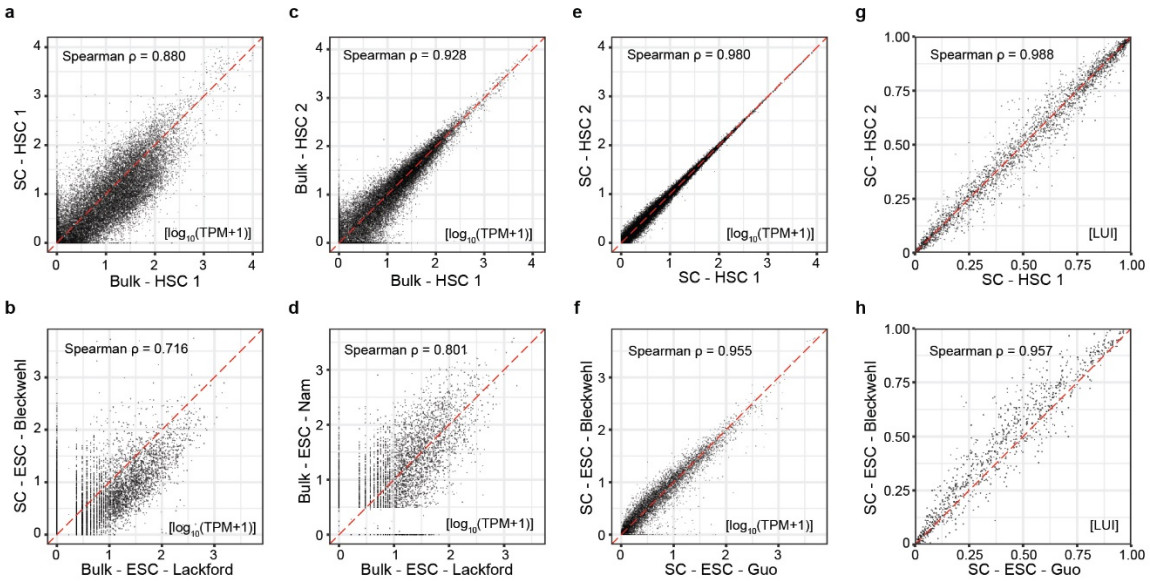

Figure 2. Accurate and precise quantification of 3′UTR isoform expression using scUTRquant.

a, Correlation of 3′UTR isoform counts obtained by bulk 3′ end sequencing (Bulk) and scUTRquant (SC) for FACS-sorted HSCs performed by different laboratories. Shown are representative samples. Correlations for additional samples are shown in Supplementary Table 2.

b, Same as a, but 3′UTR isoform counts were obtained from ESCs.

c, Correlation of 3′UTR isoform counts obtained by bulk 3′ end sequencing from biological replicate samples of FACS-sorted HSCs performed by the same laboratory. Correlations for additional samples are shown in Supplementary Table 2.

d, Same as **c**, but biological replicate samples from ESCs are shown that were generated from two different laboratories.

e, Same as **c**, but biological replicate samples were obtained by scRNA-seq.

f, Same as **d**, but biological replicate samples were obtained by scRNA-seq.

g, Same as **e**, but instead of 3′UTR isoform counts LUI values are shown.

h, Same as **f**, but instead of 3′UTR isoform counts LUI values are shown.

Quantification of 3′UTR isoform expression using scUTRquant is highly precise

Next, we assessed the reproducibility of 3′UTR transcript counts in biological replicates. We observed a substantially stronger correlation among biological replicate samples when the transcript counts were obtained from scRNA-seq samples as when they were obtained from bulk 3′ end sequencing methods (Fig. 2c-f, Supplementary Table 2). This was true when the

replicates were performed by the same laboratory or by different laboratories [26-28,50-53]. Strikingly, scRNA-seq derived 3′UTR isoform counts or isoform usage of biological replicates showed Spearman's ρ= 0.98 and ρ= 0.96 when performed by the same or by different laboratories, respectively (Fig. 2e-h, Supplementary Table 2) [50-53]. These results demonstrate a substantially lower amount of technical variation in 3′UTR isoform expression estimates when measured by scRNA-seq compared with the bulk 3′ end sequencing methods analyzed here. Because of the high accuracy and the unprecedented precision, our data suggest that 3′UTR isoform quantification from scRNA-seq data has the potential to become the new standard.

Identification of multi-UTR genes across 120 mouse cell types

After having established that scUTRquant is accurate and precise, we applied it to four scRNA-seq datasets containing 120 mouse cell types derived from embryonic stem cells, bone marrow, and most major organs, including brain [50-52,54,55]. The wide range of cell types allowed us to comprehensively identify genes that encode identical proteins but generate alternative 3′UTRs. For 3′UTR isoform usage estimation, we pooled all the cells assigned to known cell types as the current scRNA-seq data are too sparse to directly compare isoform usage across real single cells.

Fansler, Figure 3

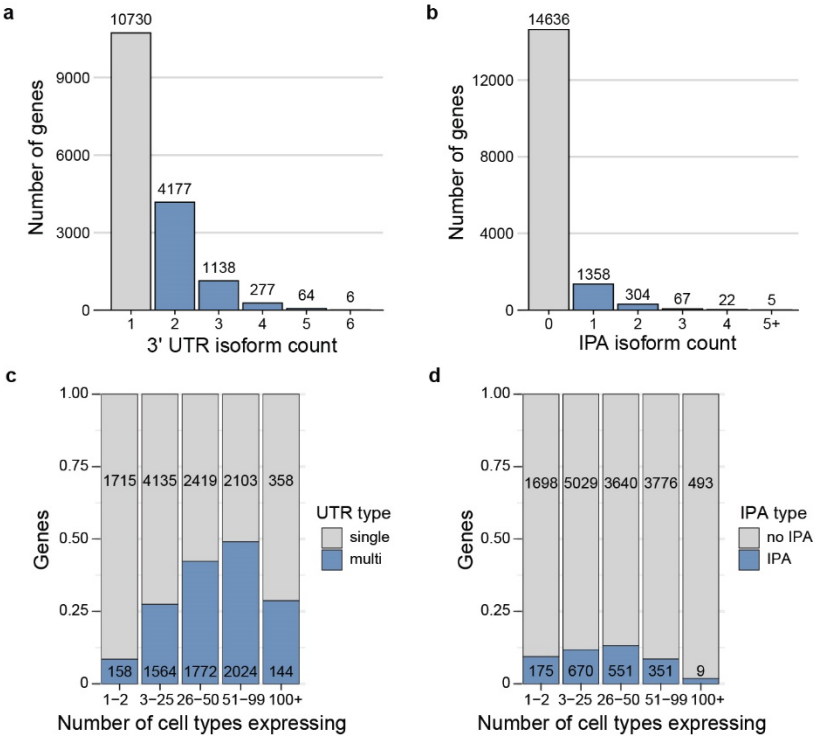

Figure 3. Characterization of multi-UTR and IPA genes across 120 mouse cell types.

a, Distribution of single- and multi-UTR genes, when using a minimum usage of 10% for each alternative 3′UTR isoform located in the terminal exon. All single- and multi-UTR genes are listed in Supplementary Table 3.

b, Same as **a**, but for IPA events. A minimum usage of 10% for each IPA isoform was required. All IPA events are listed in Supplementary Table 3.

c, Distribution of multi-UTR genes based on the number of cell types a gene is expressed in.

d, Same as **c**, but for IPA events.

In total, we detected 16,392 expressed protein coding genes (Fig. 3a). We classified a gene as multi-UTR gene if it contained at least two cleavage sites in the last exon with each of them containing a minimum of 10% of all UMI counts in the last exon in at least one cell type. We identified 5,662 multi-UTR genes (Fig. 3a, Supplementary Table 3). When instead using 5% as the minimum usage cut-off, we identified 6,395 multi-UTR genes (Supplementary Fig. 3a). The analysis of IPA isoforms using the same cut-offs detected 1,756 and 1,998 IPA isoforms, respectively (Fig. 3b, Supplementary Fig. 3b, Supplementary Table 3). For alternative 3′UTR isoforms, we observed that genes expressed in a highly cell type-specific manner had a significantly lower fraction of multi-UTR genes (Fig. 3c, χ^2 = 636, p-value < 10^{-16}). For IPA

isoforms, we observed that ubiquitously expressed genes are significantly underrepresented among genes with IPA (Fig. 3d, $\chi^2 = 42$, p-value < 10^{-11}).

scUTRboot identifies cell type-specific differences in 3′UTR isoform usage

To quantify uncertainty about mean 3′UTR isoform usage in a set of cells from a given cell type, and to identify statistically significant differences in usage between two or more cell types, we applied bootstrap-based statistical procedures, which we have collected together as an R package called scUTRboot (see methods). Specifically, we use bootstrapping to estimate confidence intervals for mean 3′UTR isoform usage. In Figures 4a and 4b, we plot the distribution of LUI for two genes across bone marrow cell types [50,51]. The width of the confidence intervals of the LUI depends on the number of cells in which a gene is expressed. Hence, when reporting point estimates or performing differential usage tests, we only calculate the LUI if a gene is expressed in at least 50 cells of a given cell type.

To test for significant differences in 3′UTR isoform usage between cell types, the mean LUI is computed for each multi-UTR gene within each cell type. A bootstrapping strategy is used to calculate a p-value for the difference in mean LUI between the cell types. Among the biological replicates shown in Supplementary Fig. 4a and 4b, we only detected a significant difference in LUI for *Lmo4* expressed in HSC (5% FDR). In contrast, scUTRboot identified significant differences in LUI between several cell types, especially in later stages of erythroblast differentiation (Fig. 4a, 4b) [50-52,54,55]. When comparing differences in 3′UTR isoform usage across all co-expressed multi-UTR genes (N = 3,153) between HSC and erythroblasts, we observed 490 significant LUI changes (Fig. 4c). In addition to requiring a significant p-value, we also require a minimum difference of 0.15 in LUI between the two cell types to consider the difference in 3′UTR isoform usage as significant (Fig. 4c). During erythroblast differentiation, we observed shortening of 3′UTRs in 436 genes (13.8%) and only observed lengthening of 3′UTRs in 64 (2.0%) genes (Fig. 4c).

The analysis of scRNA-seq data allows us to measure 3′UTR isoform expression in rare cell types. Therefore, in addition to comparing the end points of differentiation pathways, we are now able to determine the LUI in intermediate cell types (Fig. 4d). Of the 4,423 multi-UTR genes coexpressed in at least two cell types along the erythroblast differentiation pathway, we found 884 genes (20.0%) with significant LUI changes (Fig. 4d). When plotting the LUI in eight cell types during erythroblast differentiation, we observed that most genes show a gradual and coordinated change in 3′UTR isoform usage during differentiation (Fig. 4d).

When performing a similar analysis for IPA isoform expression during erythroblast differentiation, we detected a significant change in IPA isoform usage in 277 genes (Fig. 4e). However, in contrast to the predominant shortening observed for 3′UTRs, we detected similar fractions of genes with increased (N = 143) or decreased (N = 134) usage of the intronic isoform, suggesting that IPA isoforms and alternative 3′UTRs in the terminal exon are regulated independently during erythroblast differentiation.

Gene expression and 3′UTR isoform usage are independent parameters of gene regulation

As scUTRquant allows us to simultaneously measure gene and 3′UTR isoform expression, we set out to better understand the relationship between them. We used pairwise Welch *t*-tests to determine the significant gene expression changes between HSC and erythroblasts and intersected them with the changes in 3′UTR isoform usage (Fig. 4c) [56]. This analysis revealed

Fansler, Figure 4

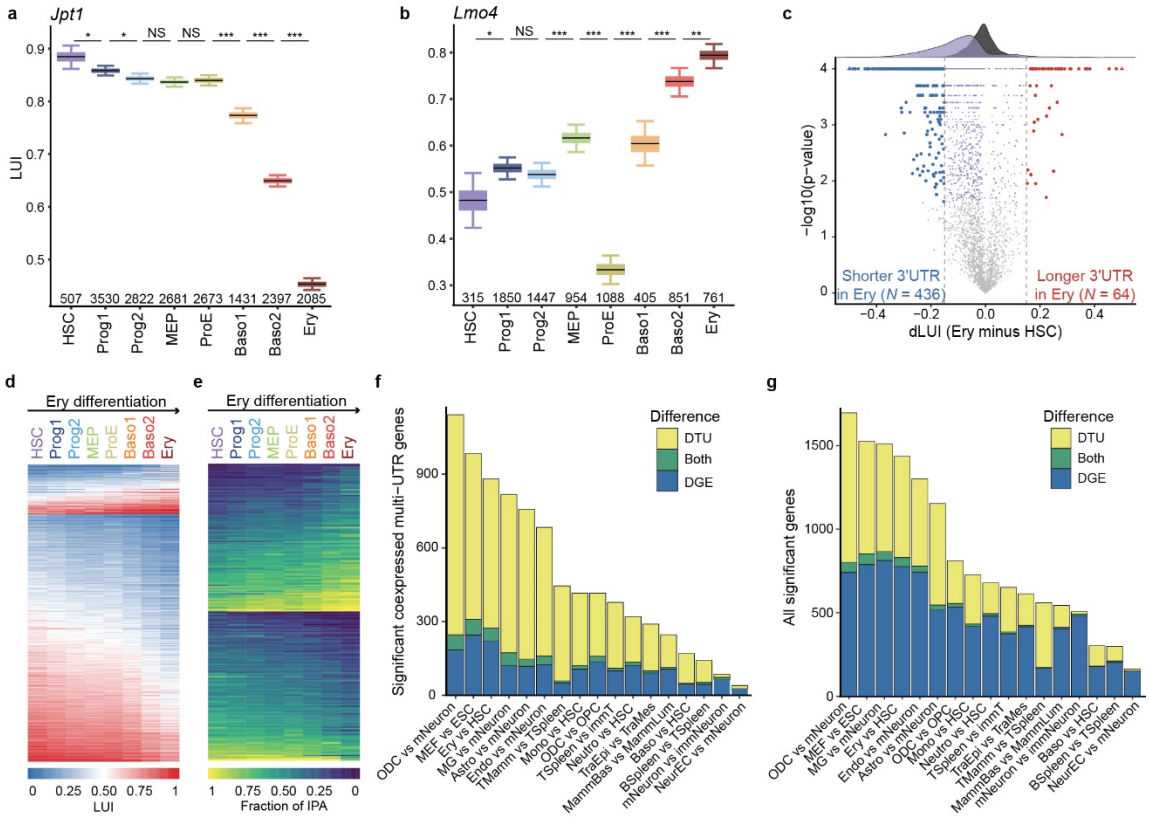

Figure 4. Gene expression and 3′UTR isoform usage are independent parameters.

a, LUI of *Jpt1* in individual cell types during erythroblast differentiation from hematopoietic stem cells (HSC). Shown is the distribution of mean LUI in each cell type obtained from bootstrapping analysis. The boxes indicate 25th and 75th percentiles and the error bars denote the range of 95% confidence interval. Prog1, progenitor cell type 1, Prog2, progenitor cell type 2, MEP, myeloid-erythroid progenitor, ProE, pro-erythroblast, Baso1, basophilic erythroblast 1, Baso2, basophilic erythroblast 2, Ery, polychromatic erythroblast. *, q-value < 0.05; **, q-value < 0.01; ***, q-value < 0.001, NS, not significant. The number of cells Jpt1 is expressed in is reported.

b, Same as **a**, but for *Lmo4.*

c, Volcano plot showing p-values and the difference in LUI (dLUI) resulting from scUTRboot's LUI bootstrap test on HSC and Ery. Purple, significant LUI changes based on q-value < 0.05, but with a difference in LUI < 0.15 (N = 1,047); light grey, genes with non-significant LUI values (q-value > 0.05).

d, Heatmap showing LUI values of each multi-UTR gene in eight cell types obtained during erythroblast differentiation. Shown are the coexpressed genes with a significant difference in scUTRboot's LUI bootstrap test for any pair of cell types. Red indicates predominant expression of the LU isoform; blue indicates predominant expression of the SU isoform.

e, Same as **d**, but shown are all genes with a significant difference in IPA isoform expression. The fraction of IPA isoform expression is color-coded.

f, Number of genes with differential gene expression (DGE) and differential 3′UTR isoform (transcript) usage (DTU) obtained from pair-wise comparison of the indicated samples. DGE was tested with Welch t-test (fold change > 1.5, q-value < 0.05); DTU was test with scUTRboot's Wasserstein's distance (WD) bootstrap test (difference in LUI > 0.15, q-value < 0.05). Corresponding values are shown in Supplementary Table 4.

g, As in **f**, but all significant genes with DTU or DGE obtained from single- and multi-UTR genes expressed in at least one cell type are shown.

that in most cases, a gene either changed its gene expression or it changed its 3′UTR isoform usage (Fig. 4f). Only 54 multi-UTR genes simultaneously changed both parameters during differentiation from HSC to erythroblasts. To examine if the two parameters are independent, we compared the observed over the expected frequency of co-regulated multi-UTR genes and found that gene expression and 3′UTR isoform usage are indeed independent parameters of

gene regulation (χ^2 = 0.14, p-value = 0.75; Supplementary Table 4). Next, we expanded this analysis to a panel of additional differentiation pathways or cell type comparisons. As changes in 3′UTR isoform usage can only be assessed for genes that are expressed in both tested cell types, we restricted the analysis to co-expressed multi-UTR genes. For 15/17 cell type comparisons changes in gene expression and changes in 3′UTR isoform usage were consistent with a null hypothesis of independence (Fig. 4f, Supplementary Fig. 4c, Supplementary Table 4). These results strongly suggest that gene and 3′UTR isoform usage contribute independent information on cell type-specific gene regulation.

However, the most significant gene expression changes are usually observed when a gene is absent from one cell type and expressed in the other. When including all genes (single- and multi-UTR) in the analysis, we observed roughly similar numbers of gene expression changes and 3′UTR isoform usage changes when comparing two cell types (Fig. 4g). This result indicates that including 3′UTR isoform usage in transcriptome analysis identifies hundreds of new genes that have previously not been known to be implicated in specific pathways or processes, as their gene expression does not significantly change during said process.

Fansler, Figure 5

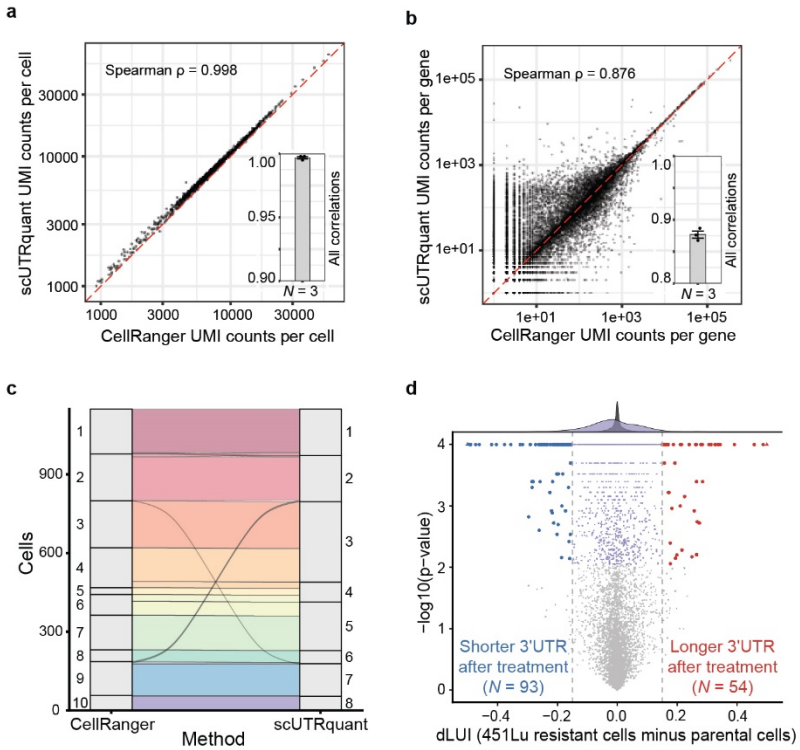

Figure 5. scUTRquant-obtained gene expression is accurate for truncated human transcriptomes obtained by txcutr.

a, Correlation of UMIs per cell obtained by scUTRquant and CellRanger for three human 10x Genomics demonstration datasets. The Spearman's ρ is shown.

b, Same as **a**, but UMIs per gene are plotted.

c, Alluvial diagram comparing Louvain clustering results derived from gene counts obtained by CellRanger and scUTRquant for the human 10x Genomics demonstration dataset PBMC_1k_v3 (ARI = 0.73).

d, Volcano plot showing p-values and the difference in LUI resulting from scUTRboot's LUI bootstrap test on 451Lu melanoma cells before and after treatment with the BRAF inhibitor PLX-4720. Purple, significant LUI changes based on q-value < 0.05, but with a difference in LUI < 0.15 (N = 1,107); light grey, genes with non-significant LUI values (q-value > 0.05). Values are reported in Supplementary Table 5.

txcutr truncates transcriptome annotations and allows 3′UTR transcript quantification in any species

We initially developed scUTRquant for the mouse transcriptome. To extend the application of scUTRquant to the transcriptomes of other species, we developed a Bioconductor package, called txcutr [57], which generates a truncated UTRome from any existing transcriptome annotation. The resulting truncated UTRome, together with a scRNA-seq dataset from the same species, can then be used as inputs for scUTRquant (Fig. 1b). To demonstrate proof-of-principle, we used txcutr to truncate (500 nucleotides) the human Ensembl Release version 93, which is used by the demonstration datasets on human data supplied by 10x Genomics. Again,

we compared gene counts per cell and per gene obtained by scUTRquant with gene counts obtained by CellRanger and observed high correlations with Spearman's ρ= 0.998 and ρ= 0.88, respectively (Fig. 5a, 5b) [49]. Also, the Louvain clustering analysis on the CellRanger-derived gene counts and the scUTRquant-derived gene counts were similar (Fig. 5c, Table 1), thus showing that both analysis pipelines can be used to quantify gene expression from scRNA-seq data with the advantage that scUTRquant also measures alternative 3′UTR isoform expression.

Finally, we used a truncated GENCODE v38 annotation and applied scUTRquant to a melanoma dataset that was generated before and after selection with the BRAF inhibitor PLX-4720 [58]. This revealed that 147 genes significantly changed their 3′UTR isoform expression (Fig. 5d, Supplementary Table 5), indicating the genes whose 3′UTR isoform changes correlate with resistance against the drug. Importantly, again, of the 213 coexpressed multi-UTR genes with differential gene expression only three of them also had significant 3′UTR isoform changes (Supplementary Table 5). This result shows that analysis of differential 3′UTR isoform usage identifies genes that are otherwise overlooked by standard differential gene expression analysis.

Discussion

The analysis of alternative 3′UTR transcripts from 3′-tagged scRNA-seq data will become a game changer for the field of APA and alternative 3′UTRs. Although the technology was developed to measure gene expression, the method is conceptually similar to 3′ end sequencing protocols, thus enabling the quantification of alternative 3′UTR isoform expression from different cell types or cell states. The abundant use of the 10x Genomics platform for gene expression analysis by researchers from diverse fields will allow re-analysis of the data, thus enabling comprehensive quantification of alternative 3′UTR isoform expression from basically any cell type, condition, or species.

In addition to being able to measure alternative 3′UTRs in any given sample, the biggest advance is the high degree of precision obtained by scRNA-seq data for quantifying 3′UTR transcripts and 3′UTR usage (Fig. 2e-h). The high degree of reproducibility observed in biological replicates performed on the same cell type by different laboratories is partially due to the use of a single experimental platform, the use of UMIs that allow the removal of PCR duplicates, and probably due to a more robust chemistry during library preparation [49]. The unprecedented reproducibility allows the integration of a large number of datasets to obtain a more comprehensive picture of alternative 3′UTR isoform expression.

In order to faithfully compare the expression of alternative 3′UTR isoforms across large numbers of datasets, we developed a predefined, high-confidence atlas of 3′ end cleavage sites for subsequent quantification of 3′UTR isoforms by scUTRquant. Most previously published computational pipelines that use scRNA-seq data as input use *de novo* peak calling on individual samples [36-40,42-44]. Without the use of a reference atlas some genes are considered single-UTR genes in some analyses but classified as multi-UTR genes in others. Our 3′ end cleavage site atlas uses a recent GENCODE annotation as its basis and adds high-confidence 3′ end cleavage sites that were obtained from pooling the sequencing data from 400,000 single cells [45]. In our cleavage site atlas, the mRNA 3′ ends were mapped from real single cells at single-nucleotide resolution. This was accomplished by using data obtained from Microwell-seq, a bead and array-based scRNA-seq method whose library preparation protocol generates longer reads with 40% of them traversing 3′ end cleavage sites [45]. The use of Microwell-seq data for the mapping of cleavage sites allowed us to overcome the biggest obstacle currently faced by other protocols that were also developed for APA quantification from 10x Genomics datasets [36-44]. Single nucleotide resolution for 3′ end cleavage sites is especially important for

downstream experimental approaches that require genetic manipulation of APA sites for the investigation of potential functions of alternative 3′UTR isoforms.

We provide here a user-friendly reusable open-access Snakemake pipeline that simultaneously quantifies gene and 3′UTR isoform expression from scRNA-seq data from any species (Fig. 1b). As input, any transcriptome annotation can be used. Therefore, the use of our augmented GENCODE annotation is optional and new releases of genome annotations can easily be incorporated. In addition to quantifying gene and 3′UTR isoform expression, we also provide a suite of statistical tools to identify genes with significant differences in 3′UTR isoform usage. As scUTRquant pseudoaligns the reads to a truncated UTRome instead of using genomic alignment, the quantification from raw sequencing data to isoform and gene counts can be performed in 5-10 mins for a 1,000-cell library on an average laptop.

scUTRquant is the first scRNA-seq analysis pipeline that was validated transcriptome-wide with respect to faithful quantification of 3′UTR isoform expression by comparing it to 3′ end sequencing protocols that are considered the current gold standard for APA quantification [26-28,33]. The validation revealed that scRNA-seq data and 3′ end sequencing protocols quantify 3′UTR isoform expression in a highly similar manner (Fig. 2a, 2b, Supplementary Table 2), thus indicating that scRNA-seq is accurate with respect to 3′UTR isoform expression measured in predefined cell types.

When we compared the clustering results for the human and mouse 10x Genomics demonstration datasets that are based on gene counts obtained by scUTRquant and CellRanger, we obtained highly similar clusters but not a perfect match (Fig. 1e, 5c, Table 1). Our analyses revealed that the largest source of discrepancy is caused by the different approaches used by kallisto and CellRanger in dealing with reads that map to more than one location in the transcriptome (Table 1, methods). Whereas CellRanger filters out multi-mapping reads, kallisto assigns these reads to multiple locations based on additional read evidence from surrounding regions [47,49]. Furthermore, scRNA-seq data contain many reads that map to internal stretches of adenosines [59]. The use of a truncated UTRome in scUTRquant removes a large fraction of reads that map to these locations. In our opinion, the strategies used by CellRanger and kallisto are both valid and it is currently unclear what approach will provide a better reflection of true gene and 3′UTR isoform expression in cells [60].

The simultaneous analysis of gene and 3′UTR isoform expression between cell types revealed that during differentiation a gene either changes its expression or it changes its 3′UTR isoform usage in most cases (Fig. 4f, 4g). This indicates that analysis of alternative 3′UTR isoforms identifies large numbers of genes that have never been implicated in specific biological processes, meaning that these analyses identify new genes that may be relevant for processes in health and disease. Our results further suggest that gene and transcript isoform expression are largely independent processes and seem to be associated with different phenotypes, which was recently suggested by the observation that genetic variants associated with gene expression are largely non-overlapping with genetic variants associated with APA [61]. These findings, together with previous reports, suggest that the primary role of cell type-specific expression of alternative 3′UTRs is not the regulation of protein abundance [1,23,25,61-63]. It is still largely unknown what the biological roles of alternative 3′UTRs are. However, the widespread quantification of alternative 3′UTRs across hundreds of cell types and conditions will have the potential to substantially increase our understanding of the regulation and function of alternative 3′UTRs.

Acknowledgements

We thank Quaid Morris for helpful input and discussions. We thank all members of the Mayr lab for helpful discussions and Sibylle Mitschka for important comments on the manuscript. This work was funded by the NIH Director's Pioneer Award (DP1-GM123454), the Pershing Square Sohn Cancer Research Alliance, and the NCI Cancer Center Support Grant (P30 CA008748).

Author contributions

M.M.F. developed and implemented all new computational tools. G.Z. performed gene expression analysis. M.M.F. and C.M. conceived the project, designed the experiments, and wrote the manuscript.

Declaration of Interests

The authors declare no competing interests.

Table 1. ARI values for Louvain clustering comparisons of mouse and human datasets.

	Mouse dataset 1 (Heart_1k_v2)			Mouse dataset 2 (Heart_1k_v3)			Mouse dataset 3 (Heart_10k_v2)		
Transcriptome annotation	CellR-clusters	scUTR-clusters	ARI	CellR-clusters	scUTR-clusters	ARI	CellR-clusters	scUTR-clusters	ARI
Default	8	9	0.80	12	10	0.84	17	15	0.77
Full-length	8	9	0.74	13	11	0.84	15	19	0.67
No multi-mapping	9	9	0.89	11	10	0.87	18	15	0.75
	Human dataset 1 (PBMC_1k_v2)			Human dataset 2 (PBMC_1k_v3)			Human dataset 3 (PBMC_10k_v3)		
Transcriptome annotation	CellR-clusters	scUTR-clusters	ARI	CellR-clusters	scUTR-clusters	ARI	CellR-clusters	scUTR-clusters	ARI
Ensembl Release v93	8	8	0.72	10	8	0.73	16	15	0.71

CellR, CellRanger; scUTR, scUTRquant

References

1 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**, 2380-2396 (2013).

2 Singh, I., Lee, S. H., Sperling, A. S., Samur, M. K., Tai, Y. T., Fulciniti, M., Munshi, N. C., Mayr, C. & Leslie, C. S. Widespread intronic polyadenylation diversifies immune cell transcriptomes. *Nature communications* **9**, 1716 (2018).

3 Lee, S. H., Singh, I., Tisdale, S., Abdel-Wahab, O., Leslie, C. S. & Mayr, C. Widespread intronic polyadenylation inactivates tumour suppressor genes in leukaemia. *Nature* **561**, 127-131 (2018).

4 Dubbury, S. J., Boutz, P. L. & Sharp, P. A. CDK12 regulates DNA repair genes by suppressing intronic polyadenylation. *Nature* **564**, 141-145 (2018).

5 Krajewska, M. *et al.* CDK12 loss in cancer cells affects DNA damage response genes through premature cleavage and polyadenylation. *Nature communications* **10**, 1757 (2019).

6 Tian, B. & Manley, J. L. Alternative polyadenylation of mRNA precursors. *Nature reviews. Molecular cell biology* **18**, 18-30 (2017).

7 Gruber, A. J. & Zavolan, M. Alternative cleavage and polyadenylation in health and disease. *Nature reviews. Genetics* **20**, 599-614 (2019).

8 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* **320**, 1643-1647 (2008).

9 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).

10 An, J. J. *et al.* Distinct role of long 3' UTR BDNF mRNA in spine morphology and synaptic plasticity in hippocampal neurons. *Cell* **134**, 175-187 (2008).

11 Lau, A. G. *et al.* Distinct 3'UTRs differentially regulate activity-dependent translation of brain-derived neurotrophic factor (BDNF). *Proc Natl Acad Sci U S A* **107**, 15945-15950 (2010).

12 Tushev, G., Glock, C., Heumuller, M., Biever, A., Jovanovic, M. & Schuman, E. M. Alternative 3' UTRs Modify the Localization, Regulatory Potential, Stability, and Plasticity of mRNAs in Neuronal Compartments. *Neuron* **98**, 495-511 e496 (2018).

13 Hafner, A. S., Donlin-Asp, P. G., Leitch, B., Herzog, E. & Schuman, E. M. Local protein synthesis is a ubiquitous feature of neuronal pre- and postsynaptic compartments. *Science* **364** (2019).

14 Ciolli Mattioli, C. *et al.* Alternative 3' UTRs direct localization of functionally diverse protein isoforms in neuronal compartments. *Nucleic Acids Res* **47**, 2560-2573 (2019).

15 Mayr, C. What Are 3' UTRs Doing? *Cold Spring Harb Perspect Biol* **11** (2019).

16 Berkovits, B. D. & Mayr, C. Alternative 3' UTRs act as scaffolds to regulate membrane protein localization. *Nature* **522**, 363-367 (2015).

17 Ma, W. & Mayr, C. A Membraneless Organelle Associated with the Endoplasmic Reticulum Enables 3'UTR-Mediated Protein-Protein Interactions. *Cell* **175**, 1492-1506 e1419 (2018).

18 Lee, S. H. & Mayr, C. Gain of Additional BIRC3 Protein Functions through 3'-UTR-Mediated Protein Complex Formation. *Mol Cell* **74**, 701-712 e709 (2019).

19 Fernandes, N. & Buchan, J. R. RPS28B mRNA acts as a scaffold promoting cis-translational interaction of proteins driving P-body assembly. *Nucleic Acids Res* **48**, 6265-6279 (2020).

20 Bae, B. *et al.* Elimination of Calm1 long 3'-UTR mRNA isoform by CRISPR-Cas9 gene editing impairs dorsal root ganglion development and hippocampal neuron activation in mice. *RNA* **26**, 1414-1430 (2020).

21 Shepard, P. J., Choi, E. A., Lu, J., Flanagan, L. A., Hertel, K. J. & Shi, Y. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA* **17**, 761-772 (2011).

22 Derti, A., Garrett-Engele, P., Macisaac, K. D., Stevens, R. C., Sriram, S., Chen, R., Rohl, C. A., Johnson, J. M. & Babak, T. A quantitative atlas of polyadenylation in five mammals. *Genome Res* **22**, 1173-1183 (2012).

23 Spies, N., Burge, C. B. & Bartel, D. P. 3' UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. *Genome Res* **23**, 2078-2090 (2013).

24 Hoque, M., Ji, Z., Zheng, D., Luo, W., Li, W., You, B., Park, J. Y., Yehia, G. & Tian, B. Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing. *Nat Methods* **10**, 133-139 (2013).

25 Gruber, A. R. *et al.* Global 3' UTR shortening has a limited effect on protein abundance in proliferating T cells. *Nature communications* **5**, 5465 (2014).

26 Lackford, B. *et al.* Fip1 regulates mRNA alternative polyadenylation to promote stem cell self-renewal. *EMBO J* **33**, 878-889 (2014).

27 Nam, J. W., Rissland, O. S., Koppstein, D., Abreu-Goodger, C., Jan, C. H., Agarwal, V., Yildirim, M. A., Rodriguez, A. & Bartel, D. P. Global analyses of the effect of different cellular contexts on microRNA targeting. *Mol Cell* **53**, 1031-1043 (2014).

28 Sommerkamp, P. *et al.* Differential Alternative Polyadenylation Landscapes Mediate Hematopoietic Stem Cell Activation and Regulate Glutamine Metabolism. *Cell stem cell* **26**, 722-738 e727 (2020).

29 Herrmann, C. J., Schmidt, R., Kanitz, A., Artimo, P., Gruber, A. J. & Zavolan, M. PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3' end sequencing. *Nucleic Acids Res* **48**, D174-D179 (2020).

30 Xia, Z., Donehower, L. A., Cooper, T. A., Neilson, J. R., Wheeler, D. A., Wagner, E. J. & Li, W. Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types. *Nature communications* **5**, 5274 (2014).

31 Gruber, A. J., Gypas, F., Riba, A., Schmidt, R. & Zavolan, M. Terminal exon characterization with TECtool reveals an abundance of cell-specific isoforms. *Nat Methods* **15**, 832-836 (2018).

32 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 biology* **19**, 45 (2018).

33 Shah, A., Mittleman, B. E., Gilad, Y. & Li, Y. I. Benchmarking sequencing methods and tools that facilitate the study of alternative polyadenylation. *Genome biology* **22**, 291 (2021).

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

35 Diag, A., Schilling, M., Klironomos, F., Ayoub, S. & Rajewsky, N. Spatiotemporal m(i)RNA Architecture and 3' UTR Regulation in the C. elegans Germline. *Developmental cell* **47**, 785-800 e788 (2018).

36 Shulman, E. D. & Elkon, R. Cell-type-specific analysis of alternative polyadenylation using single-cell transcriptomics data. *Nucleic Acids Res* **47**, 10027-10039 (2019).

37 Patrick, R., Humphreys, D. T., Janbandhu, V., Oshlack, A., Ho, J. W. K., Harvey, R. P. & Lo, K. K. Sierra: discovery of differential transcript usage from polyA-captured single-cell RNA-seq data. *Genome biology* **21**, 167 (2020).

38 Li, W. V., Zheng, D., Wang, R. & Tian, B. MAAPER: model-based analysis of alternative polyadenylation using 3' end-linked reads. *Genome biology* **22**, 222 (2021).

39 Li, G. W., Nan, F., Yuan, G. H., Liu, C. X., Liu, X., Chen, L. L., Tian, B. & Yang, L. SCAPTURE: a deep learning-embedded pipeline that captures polyadenylation information from 3' tag-based RNA-seq of single cells. *Genome biology* **22**, 221 (2021).

40 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).

41 Agarwal, V., Lopez-Darwin, S., Kelley, D. R. & Shendure, J. The landscape of alternative polyadenylation in single cells of the developing mouse embryo. *Nature communications* **12**, 5101 (2021).

42 Zhu, S., Lian, Q., Ye, W., Qin, W., Wu, Z., Ji, G. & Wu, X. scAPAdb: a comprehensive database of alternative polyadenylation at single-cell resolution. *Nucleic Acids Res* (2021).

43 Göpferich, M. *et al.* Single cell 3'UTR analysis identifies changes in alternative polyadenylation throughout neuronal differentiation and in autism. *bioRxiv*, 2020.2008.2012.247627 (2020).

44 Burri, D. & Zavolan, M. Shortening of 3' UTRs in most cell types composing tumor tissues implicates alternative polyadenylation in protein metabolism. *bioRxiv*, 2021.2006.2030.450496 (2021).

45 Han, X. *et al.* Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **173**, 1307 (2018).

46 Sheppard, S., Lawson, N. D. & Zhu, L. J. Accurate identification of polyadenylation sites from 3' end deep sequencing using a naive Bayes classifier. *Bioinformatics* **29**, 2564-2571 (2013).

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

48 Melsted, P. *et al.* Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nat Biotechnol* **39**, 813-818 (2021).

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

50 Dahlin, J. S. *et al.* A single-cell hematopoietic landscape resolves 8 lineage trajectories and defects in Kit mutant mice. *Blood* **131**, e1-e11 (2018).

51 Wolf, F. A., Hamey, F. K., Plass, M., Solana, J., Dahlin, J. S., Gottgens, B., Rajewsky, N., Simon, L. & Theis, F. J. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome biology* **20**, 59 (2019).

52 Guo, L. *et al.* Resolving Cell Fate Decisions during Somatic Cell Reprogramming by Single-Cell RNA-Seq. *Mol Cell* **73**, 815-829 e817 (2019).

53 Bleckwehl, T. *et al.* Enhancer priming by H3K4 methylation safeguards germline competence. *bioRxiv*, 2020.2007.2007.192427 (2020).

54 Tabula Muris, C. *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367-372 (2018).

55 Ximerakis, M. *et al.* Single-cell transcriptomic profiling of the aging mouse brain. *Nature neuroscience* **22**, 1696-1708 (2019).

56 Soneson, C. & Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nat Methods* **15**, 255-261 (2018).

57 txcutr: Transcriptome CUTteR. R package version 0.99.1. (2021).

58 Ho, Y. J., Anaparthy, N., Molik, D., Mathew, G., Aicher, T., Patel, A., Hicks, J. & Hammell, M. G. Single-cell RNA-seq analysis identifies markers of resistance to targeted BRAF inhibitors in melanoma cell populations. *Genome Res* **28**, 1353-1363 (2018).

59 Genomics, x. Vol. CG000376_RevA.

60 Vieth, B., Parekh, S., Ziegenhain, C., Enard, W. & Hellmann, I. A systematic evaluation of single cell RNA-seq analysis pipelines. *Nature communications* **10**, 4667 (2019).
61 Li, L. *et al.* An atlas of alternative polyadenylation quantitative trait loci contributing to complex trait and disease heritability. *Nat Genet* (2021).
62 Brumbaugh, J. *et al.* Nudt21 Controls Cell Fate by Connecting Alternative Polyadenylation to Chromatin Signaling. *Cell* **172**, 629-631 (2018).
63 Nanavaty, V. *et al.* DNA Methylation Regulates Alternative Polyadenylation via CTCF and the Cohesin Complex. *Mol Cell* **78**, 752-764 e756 (2020).
64 Amezquita, R. A. *et al.* Orchestrating single-cell analysis with Bioconductor. *Nat Methods* **17**, 137-145 (2020).
65 Lun, A. T., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research* **5**, 2122 (2016).
66 Aricode: Efficient Computations of Standard Clustering Comparison Measures (R package version 1.0, 2020).
67 Gautier, E. F. *et al.* Comprehensive Proteomic Analysis of Human Erythropoiesis. *Cell reports* **16**, 1470-1484 (2016).
68 Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome biology* **16**, 278 (2015).

Methods

mRNA 3′ end cleavage site identification from scRNA-seq data

FASTQ files for adult mouse Microwell-seq data of the Mouse Cell Atlas [45] were downloaded and then assembled using PEAR v0.9.6 with settings `-n 75 -p 0.0001`. Cell and UMI barcodes were extracted from assembled reads and placed into read headers using umi_tools v0.5.3; remaining poly-T regions at the 5′ end of assembled reads were trimmed using cutadapt v1.16, retaining only sequences with minimum length of 21 nucleotides (nts). Reads were aligned to the mm10 genome with HISAT v2.1.0. Per sample strand-specific coverage at the 5′ ends of aligned reads was computed using the `genomecov -dz -5` command of BEDTools v2.26.0; all sample coverages per strand were subsequently merged with GNU′s datamash v1.3. All entries within 3 nt radius were merged to the local mode and merge sites with at least 200 reads were classified as cleavage site candidates.

Cleavage site filtering

Candidate cleavage sites were intersected with 40 nt intervals centered at 3′ ends of GENCODE vM21 protein-coding transcripts with positively identified 3′ ends (no tag `mRNA_end_NF`). Intersecting sites (*N* = 31,196) were classified as “validated”; non-intersecting sites were subsequently intersected with 40 nt intervals centered at cluster centers in the PolyASite v1.0 mm10 Atlas supported by 3 or more experiments [29]. Intersecting sites (*N* = 25,361) were classified as “supported”; non-intersecting sites were filtered through cleanUpdTSeq v1.18.0 with maximum posterior probability of 0.0001 of being an internal priming site [46]. Passing sites (*N* = 9,214) were classified as “likely”. The union of “supported” and “likely” cleavage sites was formed and each site was annotated according to the GENCODE vM21 annotation with one of the ordered labels: “three_prime_UTR”, “five_prime_UTR”, “exon”, “intron”, “extended_five_prime_UTR”, “extended_three_prime_UTR”, or “intergenic”, where the existing 5′ ends of transcripts were extended 1 kb upstream and existing 3′ ends of transcripts were extended 5 kb downstream.

Transcriptome augmentation and truncation

The GENCODE vM21 annotation was filtered for protein-coding transcripts with known 3′ ends. Cleavage sites with a “three_prime_UTR” label were intersected with these transcripts and new transcript versions ending at the cleavage sites were generated (“upstream”). All protein-coding transcripts with known 3′ ends were extended by 5 kb downstream, intersected with the “extended_three_prime_UTR” set of cleavage sites, and new transcript versions ending at the cleavage sites were generated (“downstream”). All transcripts (GENCODE, upstream, downstream) were truncated to include 500 nts from their 3′ end. Truncated transcripts with fewer than 50 nts difference were reduced to a single representative copy, with prioritization for downstream sites. The collection of remaining truncated transcripts was exported to GTF and the corresponding sequences to FASTA. This resulted in augmenting the GENCODE annotation with 9,487 additional 3′UTR isoforms belonging to 5,931 protein-coding genes.

Empirical distributions of 10x Genomics peak width

A set of 56 peaks located at the 3′ ends of transcripts was manually curated by examining the genomic alignments of 10x Genomics Chromium v2 samples from the Tabula Muris dataset [54]. Peaks were selected for absence of splice sites, potential internal priming sites (A-rich regions), and nearby alternative cleavage sites in the immediate 800 nts upstream of the annotated cleavage site. The coverage of 5′ ends of reads was extracted with the `bedtools genomecov -5` command for each sample from the Tabula Muris dataset and the distance from the annotated cleavage site of the corresponding transcript was computed. For each gene-sample combination, the 95th percentile for distance from the 3′ end was computed. Additionally, 95th

percentiles were computed for each gene and sample, aggregating across samples and genes, respectively (Supplementary Fig. 1b-d).

Kallisto transcript quantification resolution

The sequence of the Ensembl transcript Rac1-201 (ENSMUST00000080537) was used as the basis for a two-isoform transcript expression simulation. The first simulated isoform ("distal") used the annotated 3′ end; the second ("proximal") was created by removing specified intervals from the 3′ end. For each round of simulation, samples of read distances from the 3′ end of each transcript were generated according to a discretized gamma distribution with mean 300 and standard deviation of 100. Reads of 100 nts were generated using the respective transcript sequences and the randomly sampled positions. The `kallisto quant` command was used to estimate transcript abundance, using the parameters `--single -l1 -s1 --fr-stranded --pseudobam` and truncated versions of the transcripts as index. Relative error for each transcript was computed using estimated and true abundances. A parameter sweep was performed with all combinations of the following parameters: (a) cleavage site distances between [50-700] with 50 nt steps; (b) truncated transcript lengths [350-600] with 50 nt steps; (c) proximal counts {50,100}; (d) distal counts {50,100}. Each parameter combination was simulated for 10 replicates. Final resolution was selected based on mean relative errors approaching zero (Supplementary Fig. 1e).

Kallisto customization and scUTRquant settings

The `kallisto bus` command of kallisto version 0.46.2 was extended to support strand-specific pseudoalignment for both FASTQ and BAM input files (ref: https://github.com/mfansler/kallisto/releases/tag/v0.46.2sq). All 10x Genomics 3′ end datasets were pseudoaligned with `kallisto bus --fr-stranded`. Cell barcodes for the corresponding technology version (v2 or v3) were used as whitelists for the `bustools correct` command. Truncated isoforms in the same gene with 3′ ends nearer than 200 nts apart (N = 7,022) were merged in the `bustools count` step.

Additional scUTRquant indices generated with txcutr

The default target for scUTRquant uses the mouse UTRome described above. Additional scUTRquant targets were created with `txcutr` v0.99.0 (functionally equivalent to the Bioconductor release v1.0.0) to generate truncated GTF annotations, FASTA sequences, and merge tables [57]. All indices used a 500 nt truncation length and a merge distance of 200 nts. In brief, all GENCODE annotations were first pre-filtered with an AWK script to remove any entries with the `mRNA_end_NF` tag (indicating unvalidated 3′ ends) and then restricted to protein-coding transcripts. The txcutr method `truncateTxome` clips all transcripts longer than the specified length, anchored at the 3′ end, intersects the truncated transcripts with the child exons of that transcript, and then redefines the genomic range of the gene to the union of all child transcripts. Transcripts that are identical after truncation are deduplicated to retain only one representative copy, which is annotated with the transcript ID of the transcript with lexicographical priority. The resulting TxDb object is then exported as a GTF and a FASTA file using txcutr's `exportGTF` and `exportFASTA` methods, respectively. Finally, a merge table is generated with txcutr's `generateMergeTable` by further truncating transcripts to the specified merge distance, anchored at the 3′ end, intersecting within the parent gene, and recording the most downstream transcript with which each intersects. Additional specification and implementation details are found in the txcutr documentation. All indices used in these results are reproducible from the Snakemake pipeline available at https://github.com/Mayrlab/txcutr-db.

CellRanger and sqUTRquant UMI count correlations

Six 10x Genomics 3′ end mouse demonstration datasets were downloaded as FASTQ files from the 10x Genomics website (`heart_1k_v2`, `heart_1k_v3`, `heart_10k_v3`, `neuron_1k_v2`, `neuron_1k_v3`, and `neuron_10k_v3`) and processed through the scUTRquant pipeline using default settings. The corresponding filtered HDF5 UMI counts from CellRanger 3.0.0 were also downloaded and loaded as SingleCellExperiment objects in R [49]. For each dataset, only cells (or genes) present in both the CellRanger and scUTRquant results were plotted and used to compute Spearman correlations.

Similarly, three 10x Genomics 3′ end human demonstration datasets (`pbmc_1k_v2`, `pbmc_1k_v3`, and `pbmc_10k_v3`) were processed using scUTRquant with the Ensembl Release 93 annotation preprocessed according to the CellRanger 3.0.0 pipeline and truncated to 500 nts using txcutr. Comparisons were performed against CellRanger UMI counts in the same manner as above.

CellRanger and sqUTRquant clustering comparisons

For each 10x Genomics dataset, the CellRanger and scUTRquant counts were filtered to common cells. Clustering was performed following Amezquita et al., (2019) [64]. In brief, size factors were computed with the `computeSumFactors` from the `scran` Bioconductor package [65], and then used to compute normalized log counts. The top 1000 high-variance genes were used to compute the first 20 principal components. Louvain clustering was performed on the cells in this reduced representation. The Adjusted RAND Index (ARI) between the CellRanger and scUTRquant clusters was computed using the `aricode` R package [66].

<u>CellRanger versus scUTRquant clustering diagnostic analyses:</u>

In order to identify sources of difference in the clustering of gene counts obtained from CellRanger and scUTRquant, two additional comparisons were made. To test whether the differences were due to exclusion of reads from internal priming sites that are excluded *a priori* by truncation, a full-length version of the UTRome annotation was used as the target in scUTRquant ("Full-length"; Table 1). Clustering was then performed following identical procedures as described above.

Another possible source of difference we considered was multimapping reads. These are excluded in the CellRanger 3.0.0 pipeline but are retained in the scUTRquant pipeline. To assess this, each kallisto equivalence class was checked for whether all transcripts overlapping in the equivalence class belonged to the same gene. When this was not true, the genes involved were marked as having multimapping reads. All such genes were then filtered out at the start of the clustering analysis, preventing their consideration in the PCA reduction step ("No multi-mapping"; Table 1).

Classification of single- and multi-UTR genes from 120 mouse cell types

Samples from embryonic stem cells (ESCs; GEO:GSM3629847-8), Tabula Muris (GEO:GSM3040890-917), bone marrow (GEO:GSM2877127-32), and brain datasets (GEO:GSM3722100-115) were quantified for transcript usage following the default settings of scUTRquant and cells were annotated with published cell type annotations [50-52,54,55]. Cell type labels for bone marrow cell types were obtained by combining publicly available transcriptome and proteome information for erythroblast differentiation [51,67]. Cells not previously annotated in published analyses were excluded.

All datasets were merged into one SingleCellExperiment object and counts were size-factor normalized using the `computeSumFactors` method from Bioconductor package `scran` [65]. UMI counts were aggregated by cell type and the percentage of isoform usage per gene was computed, excluding isoforms whose 3′ ends were located within a GENCODE-annotated intron

of the corresponding gene. For each gene, the number of isoforms with at least 10% usage in at least one cell type were counted. Genes with two or more such isoforms were classified as multi-UTR genes; otherwise, they were classified as single-UTR genes. To identify intronic polyadenylation (IPA) genes, all mRNA 3′ ends of a transcription unit were included. IPA isoforms were counted if they contained at least 10% of reads of a gene in at least one cell type. The Snakemake pipeline for classification is available at https://github.com/Mayrlab/atlas-mm.

Comparison of 3′UTR isoform counts obtained by scUTRquant with bulk 3′ end sequencing methods

FACS-sorted hematopoietic stem cells (HSCs):

FASTQ files from FACS-sorted HSC samples of Sommerkamp et al., (2020) were downloaded from ArrayExpress (E-MTAB-7391) and 3′UTR isoforms were quantified by pseudoalignment of R2 using the UTRome annotation and `kallisto quant` [28]. This was compared to scRNA-seq from Dahlin et al., (2018) using the annotations of Wolf et al., (2019) to filter for early HSCs (clusters 0 and 1) [50,51]. The scRNA-seq UMI counts were aggregated by sample, and transcript per million (TPM) per sample was computed by normalizing to UMIs per million. LUI values were computed for all multi-UTR genes expressing exactly two 3′UTR isoforms present in the last exon of all samples.

ESC datasets:

3′UTR isoform UMI counts were quantified for scRNA-seq ESC datasets using the scUTRquant pipeline with default settings and the mouse UTRome [52,53]. TPM values and cleavage site locations for bulk ESC 3′ end sequencing datasets [26,27] were obtained from the PolyASite v2.0 database [29]. Bulk cleavage sites were intersected with the UTRome annotation using a 50 nt interval around PolyASite cluster centers. The corresponding TPM values for all PolyASite clusters intersecting a given UTRome transcript were summed to yield a translation of PolyASite quantifications to UTRome quantifications. The scRNA-seq data was summarized to TPM per sample by aggregating counts across all cells in each sample and normalizing to UMIs per million. LUI values were computed for all multi-UTR genes expressing exactly two 3′UTR isoforms present in the last exon of all samples.

Bootstrap mean LUI estimates

For each cell type (Fig. 4a, 4b) or library (Supplementary Fig. 4a, 4b), 2000 bootstrap samples were generated by resampling with replacement from the pool of all cells with that cell type or library annotation. The LUI was computed for each bootstrap sample. Percentile statistics were then calculated for these values across the bootstrap samples to determine the confidence interval on the mean LUI.

Two-sample bootstrap test with scUTRboot

The R package `scutrboot` implements two-sample hypothesis testing with a bootstrap strategy for estimating p-values. The `twoSampleTest` function provides two general modes of tests based on the statistic computed across the samples: either a *Usage Index (UI)* or a *Wasserstein Distance (WD)*, also called the Earth Mover′s Distance.

For the UI statistic, users provide a feature (`featureIndex`), such as short 3′UTR isoform (SU), long 3′UTR isoform (LU), or intronic polyadenylation isoform (IPA), for each gene. The UI statistic per gene is computed as the difference in the fraction of usage of this isoform in the gene across the two sets of cells. This characterizes the difference across sets of cells for a *single feature*.

Alternatively, the WD statistic per gene is computed as half the total difference in all isoform usages in the gene across the two sets of cells. When a gene has exactly two isoforms, the UI

and WD statistics are identical in magnitude. However, the WD statistic is sensitive to changes in any isoform, including cases when there are more than two isoforms.

For either statistic, p-values per gene are estimated using bootstrap resampling under the null hypothesis that the two samples of cells came from identically distributed populations. Specifically, the union of the two samples of cells is used to sample with replacement sets of cells of the same size as the original samples. For each bootstrap sample, the statistic is computed for each gene and the p-value is estimated as the fraction of bootstrap statistics as extreme or greater than the observed statistic, with a pseudocount of 1 included to provide a conservative upper bound for rare events.

scUTRboot includes a `minCellsPerGene` option to exclude genes that are not sufficiently coexpressed in the samples to compare with confidence. When this is set, bootstrap samples that do not satisfy this minimum are discarded and the p-value will only be computed from the retained samples. The number of bootstraps samples used to estimate the p-value is included in the test results.

Pairwise two-sample bootstrap tests on HSC to erythroblasts differentiation trajectory:

All tests were performed on size-factor normalized UMI counts. scUTRboot was used to perform a two-sample LUI test (`featureIndex="is_distal"`) and IPA (`featureIndex="is_ipa"`) tests on all pairs of all cell types (8 cell types, 28 unique pairs) along the HSC to erythroblast trajectory from the bone marrow dataset [50,51] using 10,000 bootstrap samples on all co-expressed genes (minimum 50 cells expressing each gene) and corrected for multiple testing using Benjamini-Hochberg procedure. Genes with at least a LUI difference > 0.15 and q-value < 0.05 were classified as significant.

Comparing differential gene expression with differential 3′UTR isoform usage

Differential gene expression was performed on pairs of cell types following Amezquita et al., (2019) [64]. In brief, gene-level UMI counts were log-normalized using size factors and a pseudocount of 1 and differential expression was tested with a Welch *t*-test [56]. All p-values were corrected using the Benjamini-Hochberg procedure and genes were classified as significant if fold-changes exceeded 1.5 in either direction and q-value < 0.05. To check against an alternate differential gene expression pipeline procedure, we converted gene-level UMI counts to Seurat objects, normalized with SCTransform, and performed pairwise tests with MAST [68]. We observed comparable numbers of significant genes when either using MAST or Welch *t*-test. We report the Welch *t*-test results.

To identify genes with differential 3′UTR isoform usage, two-sample WD tests were performed on all cell type pairs shown in Fig. 4f using scUTRboot on size-factor normalized UMI counts. All p-values were corrected using the Benjamini-Hochberg procedure and genes were classified as significant if WD > 0.15 and q-value < 0.05.

To test if gene expression and 3′UTR isoform usage are independent, for each comparison, all coexpressed multi-UTR genes were classified as either non-significant, DGE only, DTU only, or both. A Chi-Square test for independence was performed on the resulting tabulation.

BRAF inhibitor-resistant melanoma data set analysis:

The 10x Genomics Chromium single cell 3′ samples (GEO: GSM2897333-4) were processed using scUTRquant with the GENCODE v38 annotation truncated to 500 nts using txcutr [58]. Cells with more than 10% of UMIs coming from mitochondrial genes and those with low transcript diversity, as measured by Shannon′s diversity index < log2(500), were excluded as low quality. UMI counts were size factor-normalized. A two-sample LUI test between 451Lu-resistant and 451Lu-parental samples was performed using scUTRboot, p-values corrected with the

Benjamini-Hochberg procedure, and genes with a LUI difference > 0.15 and q-value < 0.05 were classified as significant. Log-normalized UMI counts at the gene-level were used to test differential gene expression with a Welch *t*-test, as above.