

# Single-cell RNA-seq using UltraMarathonRT expands the known transcriptome

Chia-Ling Chou<sup>1#</sup>, Anastasiya Grinko<sup>1#</sup>, Li-Tao Guo<sup>2,3</sup>, Alexander M. Leipold<sup>1,4</sup>, Teresa Rummel<sup>5</sup>, Florian Erhard<sup>5</sup>, Anna Marie Pyle<sup>2,6,7</sup>, Antoine-Emmanuel Saliba<sup>1,4,8\*</sup>

<sup>1</sup>Helmholtz Institute for RNA-based Infection Research (HIRI), Helmholtz Centre for Infection Research (HZI), Würzburg, Germany.

<sup>2</sup>Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, Connecticut 06520, USA

<sup>3</sup>RNAConnect, Branford, Connecticut 06405, USA

<sup>4</sup>Institute of Molecular Infection Biology (IMIB), University of Würzburg, Würzburg, Germany.

<sup>5</sup>Faculty for Informatics and Data Science, University of Regensburg, Regensburg, Germany.

<sup>6</sup>Department of Chemistry, Yale University, New Haven, Connecticut 06520, USA.

<sup>7</sup>Howard Hughes Medical Institute, Chevy Chase, Maryland 20815, USA.

<sup>8</sup>Faculty of Medicine, University of Würzburg, Würzburg, Germany.

# Contributed equally

\* Corresponding author

Correspondance: [emmanuel.saliba@helmholtz-hiri.de](mailto:emmanuel.saliba@helmholtz-hiri.de)

**Keywords:** UltraMarathonRT; reverse transcriptase; template switching; single-cell RNA-seq; metabolic labeling; scSLAM-seq

## Abstract

The ability to map messenger RNA (mRNA) molecules from individual cells using next-generation sequencing technologies, known as single-cell RNA-seq (scRNA-seq), is transforming biology by redefining cellular identities with unmatched detail. However, all current protocols depend on copying RNA into complementary DNA with a single reverse transcriptase (RT) derived from murine leukemia virus, which is an RT enzyme known for low processivity and limited ability to unfold complex RNA structures. Here, for the first time, we introduce a group II intron reverse transcriptase, UltraMarathonRT (uMRT), to perform scRNA-seq. We demonstrate that this enzyme reveals an unexpected transcriptomics landscape by capturing additional genes and other genomic features that conventional RTs miss. We also combined uMRT with metabolic RNA labeling, nucleoside conversion and scRNA-seq to explore genome-wide transcriptome dynamics at the single-cell level. Overall, we establish uMRT as a transformative biotechnological tool for single-cell transcriptomics.

## Introduction

Single-cell (sc)RNA-seq, defined as genome-wide transcriptomics technology at the single-cell level, is significantly transforming cell taxonomy (Tanay and Regev, 2017) and uncovering the wide variety of cell states present throughout an organism (Regev *et al.*, 2017). Essentially, the technology's strength lies in its dimensionality, enabling the measurement of expression levels for thousands of genes simultaneously (Mereu *et al.*, 2020), which allows for a data-driven definition of cell identities (Tanay and Regev, 2017). The technology aims to enumerate all RNA molecules in a cell in an unbiased manner, providing a unique identity profile for each cell. Experimentally, scRNA-seq protocols emerged by adapting and improving template-switching-based transcriptomics methods (Ramsköld *et al.*, 2012; Saliba *et al.*, 2014). Such protocols typically follow a three-step process: (i) poly-adenylated messenger RNA molecules are reverse-transcribed into first-strand cDNA; (ii) a primer sequence is added to the cDNA using a template-switching mechanism, allowing the generation of the second strand; and (iii) cDNA is amplified to reach detection sensitivity before sequencing (Saliba *et al.*, 2014). Improvements in the workflow have focused on increasing the number of cells analyzed (Svensson, Vento-Tormo and Teichmann, 2018), and enhancing assay sensitivity (Mereu *et al.*, 2020). Additionally, commercial platforms have been developed and deployed to streamline access to the technology (Zheng *et al.*, 2017). However, if a cell can be viewed as a “bag of RNA” (Quake, 2021), a natural question arises: do we accurately capture the true picture of RNA molecules in cells?

Central to all scRNA-seq protocols is the ability to reverse transcribe RNA into cDNA using reverse transcriptase (RTs), assuming this step faithfully reflects the RNA census. Published protocols have relied on retroviral RTs derived from a single enzyme originating from Moloney murine leukemia virus (MMLV), which has the power to perform template-switching (Kulpa, Topping and Telesnitsky, 1997; Matz *et al.*, 1999). It is produced and marketed in recombinant form with popular names like SuperScript, Maxima, SmartScribe (Zhu *et al.*, 2001; Zucha *et al.*, 2020; Martín-Alonso, Frutos-Beltrán and Menéndez-Arias, 2021; Verwilt, Mestdagh and Vandesompele, 2023). However, extensive research has shown that these MMLV variants introduce biases in RNA sequencing due to low processivity and fidelity (Mohr *et al.*, 2013; Guo *et al.*, 2022; Verwilt, Mestdagh and Vandesompele, 2023). RT families are not restricted to retroviral RT variants, instead spanning across the tree of life with diverse functions (González-Delgado *et al.*, 2021; Martín-Alonso, Frutos-Beltrán and Menéndez-Arias, 2021). Therefore, uncovering the full complexity of the transcriptome requires the use of alternative RTs. Recently, RTs from mobile group II introns have been developed as biotechnology tools for bulk RNA-seq workflows, known as TGIRT, InduroRT, and MarathonRT, exhibiting

superior properties compared to MMLV enzymes (Mohr *et al.*, 2013; Zhao, Liu, *et al.*, 2018). Harnessing these enzymes has involved designing the entire RNA-seq process de novo, with the main challenge being to introduce sequence handles for second-strand synthesis and cDNA amplification. Recently, MarathonRT, and its improved version UltraMarathonRT (uMRT), have been shown to perform RNA-seq via a template-switching mechanism, revealing previously unseen transcripts and generating longer cDNA than established MMLV RT (Guo *et al.*, 2024). Given their many advantages, we aimed to incorporate a Group II intron-based RT into the scRNA-seq workflow.

Here, we present a comprehensive scRNA-seq method based on uMRT that achieves picogram (pg) sensitivity. We demonstrate that it is suitable for performing efficient scRNA-seq and we compare it to an established protocol, showing that it uncovers new, unexpected transcriptomic space at the single-cell level. Finally, we also establish that our workflow is compatible with scSLAM-seq, encompassing RNA metabolic labeling, nucleotide conversion, and scRNA-seq to measure RNA dynamics of thousands of genes. Overall, our work represents the first integration of a new class of ultraproccessive RT into a scRNA-seq workflow.

## Results

### Picogram-level RNA-seq libraries can be generated using uMRT

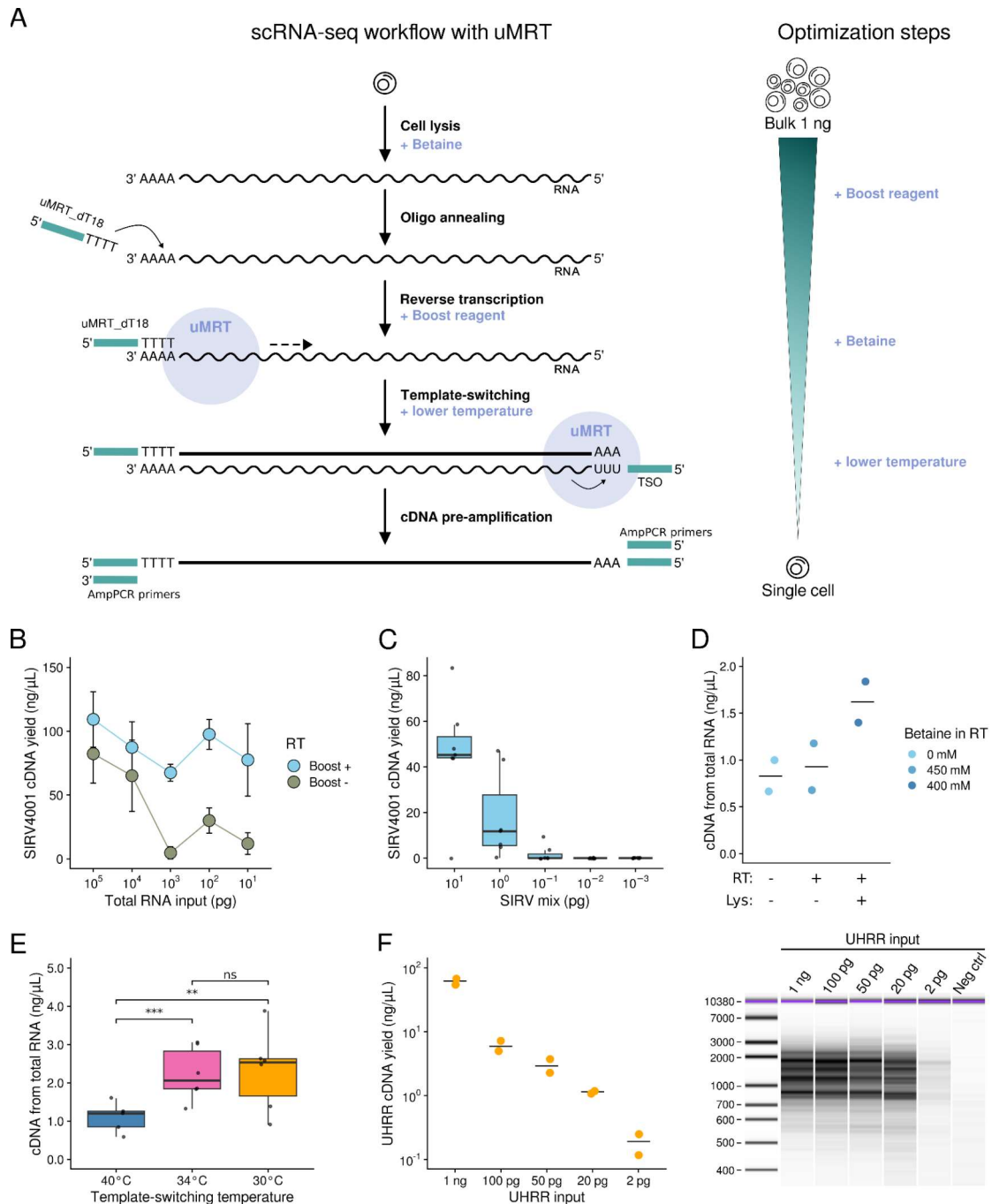
Our first objective was to demonstrate that we can reach RNA picogram level of sensitivity using a uMRT-based RNA-seq workflow given that previously published protocols were not designed to generate cDNA on picogram levels of RNA (Guo *et al.*, 2024). Therefore, we systematically modified the experimental conditions at each step of the uMRT-based RNA-seq workflow using purified RNA mixes as input (**Figure 1A**).

First, we optimized reverse transcription conditions on a single RNA target. For this purpose, we chose SIRV4001, a 4-kilobase-long synthetic transcript contained in Lexogen's Spike-In RNA Variants mix 4 (SIRV-Set 4), to serve as a mimic for a polyadenylated full-length transcript. To determine uMRT's sensitivity, we spiked  $7 \times 10^4$  copies of SIRV4001 RNA into decreasing amounts of total mouse brain RNA, with a total RNA input ranging from 100,000 pg down to 10 pg. Reverse transcribed full-length SIRV4001 RNA was then amplified using target-specific primers (see Methods). At the same time, we tested whether the RNAConnect buffer component known as 'Boost' could enhance the capture of low RNA amounts. This buffer was developed to increase the template-binding efficiency of uMRT, which is especially crucial when the template is scarce as in the case of single cells. Indeed, in presence of Boost,

uMRT's capacity to capture and reverse transcribe 150 fg of SIRV4001 RNA remained stable, even with decreasing amounts of total RNA (**Figure 1B**), demonstrating that Boost enables uMRT to capture low-input RNA and completely eliminating the need for carrier RNA. This is further highlighted by the fact that Boost enables uMRT reverse transcription of as little as 0.1 pg SIRV mix and  $7 \times 10^2$  copies of SIRV4001 RNA, without any carrier RNA in the environment (**Figure 1C**).

After enhancing uMRT's sensitivity for target amplicons, we sought to optimize transcriptome-wide capture by harnessing the template-switching activity of uMRT (Carninci *et al.*, 1998; Guo *et al.*, 2024). To that end, we used two commercially available RNA extract mixes as input, one from mouse brain tissue and a mix of different cell lines called UHRR (Methods). Our reverse transcription reaction was launched using an oligodT primer and it followed the same steps as described earlier (Guo *et al.*, 2024).

Previously, betaine has been reported to increase the yield of MMLV-based Smart-seq2 protocol (Picelli *et al.*, 2013; Hahaut *et al.*, 2022). We therefore tested different concentrations of betaine at different steps of our protocol. When added to the reverse transcription step at 450 mM, betaine increased the cDNA yield only minimally, whereas the effect was more enhanced (1.5-fold) when added in the lysis buffer at 1600 mM and remaining in the reverse transcription reaction at 400 mM (**Figure 1D**). While betaine acts as a molecular crowding agent (Picelli *et al.*, 2013; Deng, Chapagain and Leng, 2023), it has also been reported to stabilize proteins (Santoro *et al.*, 1992). We hypothesized that greater protein stability would further increase cDNA library yield. In our hands, known protein stabilizers such as PEG8000 (Kramer *et al.*, 2012) and trehalose (Chan *et al.*, 1980; Carninci *et al.*, 1998; Guo *et al.*, 2024) did not show any convincing effect (**Supplementary Figure 1C, D**), however, lowering the temperature during template-switching increased the cDNA yield two-fold (Student's *t*-test,  $p$ -value = 0.0096 between 40°C and 34°C,  $p$ -value = 0.037 between 40°C and 30°C), resulting in sufficient cDNA amounts for library preparation (**Figure 1E**). Indeed, our optimization efforts enabled the generation of well-defined cDNA from as little as 2 pg of extracted total RNA (**Figure 1F**), giving us sufficient confidence to perform further optimizations directly in single cells.



**Figure 1: Generation of cDNA libraries with uMRT down to picogram-level of RNA input.**

**(A)** Step-by-step overview of our uMRT-based (sc)RNA-seq workflow with optimization steps highlighted in lavender.

**(B)** Amplicon yield of the 4 kb synthetic SIRV4001 RNA originating from Lexogen's SIRV-Set 4 mix in the presence or absence of Boost buffer. Total RNA input consisted of 10 pg of SIRV-Set 4 mix in each sample, topped with total mouse brain RNA to the indicated amount. Points indicate the mean (n=3); error bars indicate standard error of the mean.

**(C)** Boxplot showing the yield of SIRV4001 amplified from SIRV-Set 4 mix as input in absence of additional total RNA (n=7).

**(D-E)** cDNA libraries generated from 30 pg of total mouse brain RNA across various conditions. **(D)** Effect of betaine on cDNA yield at the lysis and/or RT step of the protocol (n=2). **(E)** Impact of the template-switching temperature on cDNA yield. Boxplots show the cDNA yields obtained at 40°C (blue), 34°C (pink) and 30°C (yellow) (n=6). Differences were calculated with Student's *t*-test.

**(F)** cDNA quantity (left) and Bioanalyzer profile (right) of libraries (n=2) obtained from Universal Human Reference RNA (UHRR) over the range of 1 ng to 2 pg, by combining all of the above optimized experimental conditions. Horizontal bar shows the mean with individual samples shown as points.

## Optimization of uMRT reaction conditions to perform scRNA-seq

With uMRT optimized to generate sufficient amounts of cDNA at low RNA inputs, we continued our optimizations in single cells. We chose as a model cell line K562, a tier-one cell line from ENCODE, which we sorted in lysis buffer-filled plates using FACS. In parallel, from the exact same cell sorting experiment, we reserved control cells for processing with the MMLV-based Smart-seq protocol available from Takara SMART-seq (Ramsköld *et al.*, 2012). As our reference, we extracted RNA from K562 cells and used 1 ng for our assays, denoted as ‘bulk’ hereafter.

Since our protocol relies on template-switching, we investigated the effects of template-switching temperature and incubation time on the sequencing metrics of our method. For this, we tested incubation at 30°C for 60 min (hereafter referred to as the 30/60 protocol) and at 37°C for 30 min (hereafter referred to as the 37/30 protocol). As previously observed in UHRR samples, we obtained sufficient amounts of cDNA from K562 cells for both protocols, with the 30/60 protocol achieving higher cDNA amounts (1.02 ng/μL ± 0.28 standard deviation (sd) for single cells, 5.44 ng/μL ± 0.16 sd for bulk) than the 37/30 protocol (0.29 ng/μL ± 0.10 sd for single cells, 2.30 ± 0.51 sd for bulk)(**Figure 2A**). The captured cDNA fragments had a size distribution between 700 and 3000 bp (**Figure 2B**), which were subjected to Illumina library preparation and sequencing at a maximized sequencing depth (11 million reads for single cells, 20 million reads for bulk).

While template-switching methods are intended to generate amplicons that span from one primer to another, one can observe uncontrolled continuation of switching templates, resulting in TSO concatemers that flood sequencing runs (Kapteyn *et al.*, 2010). Here, we find that for uMRT, the TSO concatemer rates remained below 3 % on the single-cell level with lower rates in the 30/60 protocol (2.09 % ± 0.64 sd) than in the 37/30 protocol (5.68 % ± 2.74 sd) (**Figure 2C**). On the bulk level, the 30/60 protocol showed TSO concatemer rates with an average of 0.25 % (sd = 0.05), and the 37/30 protocol showed average rates of 0.09 % (sd = 0.06). After read trimming and filtering out low quality reads, the most of the reads for every protocol could be retained for the bulk protocols (97.86 % ± 0.6 sd for 30/60 and 98.19 % ± 0.08 sd for 37/30) and also on the single-cell level (94.19 % ± 1.18 sd for 30/60 and 90.12 % ± 2.25 sd for 37/30, respectively). Of those high-quality reads, the majority mapped

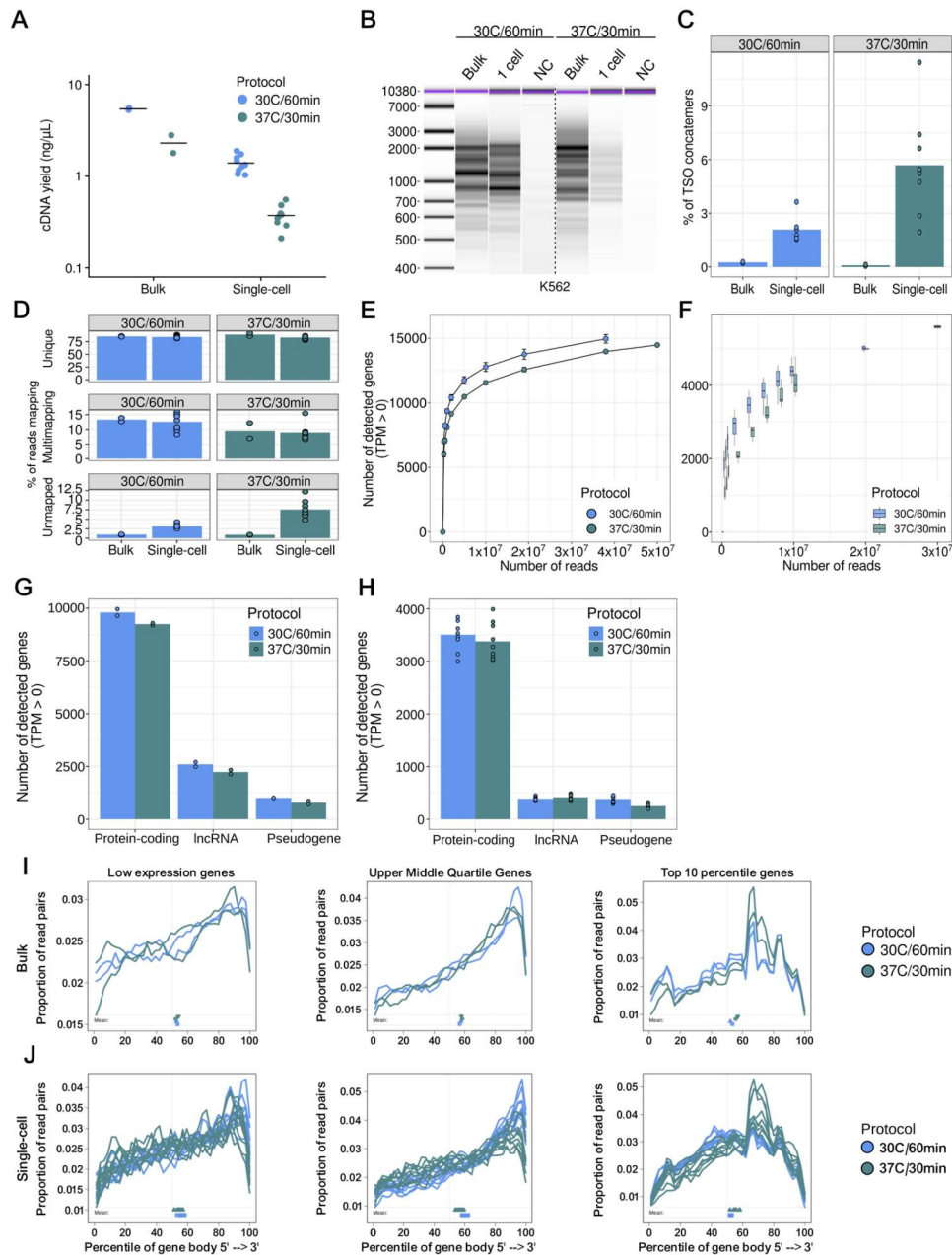
uniquely to the GENCODE 47 human reference set for both protocols (**Supplementary Table 1, Figure 2D**). On the bulk level, the 37/30 protocol showed a slightly higher percentage of uniquely mapped reads ( $88.97\% \pm 3.87$  sd) than the 30/60 protocol ( $85.4\% \pm 1.07$  sd), whereas on the single-cell level, the 30/60 protocol showed slightly higher percentages of uniquely mapped reads ( $84.19\% \pm 2.75$  sd) in comparison to the 37/30 protocol ( $83.27\% \pm 3.87$  sd). To assess the completeness of transcript detection, we performed *in silico* downsampling to obtain sequencing saturation curves, simulating sequencing depths of 0.25 to 50 million reads. As expected, the number of detected transcripts increased with increasing sequencing depth. In the bulk samples, plateauing began at around 20 million reads, with the 30/60 protocol showing a steeper curve than the 37/30 protocol (**Figure 2E**). This trend was also observed at the single-cell level (**Figure 2F**). However, sequencing saturation was not reached yet at 10 million reads, indicating the need to sequence deeper for full completeness. To evaluate the gene detection sensitivity of both protocols, we compared the number of detected genes (transcripts per million (TPM)  $\geq 0$ ) between both uMRT protocols on the bulk and single-cell level. Accounting for effects of sequencing depth, we downsampled the number of reads to 19 million reads on the bulk level and 10 million reads on the single-cell level. Additionally, we also elucidated detected genes across the three most abundant biotypes, protein-coding, lncRNA and processed pseudogenes (according to GENCODE 47).

Overall, the 30/60 uMRT protocol detected more genes on average with an average total of  $13,657 \pm 385.4$  sd genes on the bulk level and  $4,350 \pm 313.2$  sd genes on the single-cell level compared to  $12,577 \pm 174.66$  sd genes and  $4,103 \pm 386.1$  sd genes for the 37/30 uMRT protocol, respectively. On the bulk level, the 30/60 protocol consistently detected more genes across each biotype ( $9,799 \pm 217.79$  sd protein-coding,  $2,589 \pm 153.44$  sd lncRNA,  $1,009 \pm 4.95$  sd processed pseudogenes) than the 37/30 protocol ( $9,244 \pm 67.88$  sd protein-coding,  $2,234 \pm 128.59$  sd lncRNA,  $786 \pm 111.72$  sd processed pseudogenes) (**Figure 2G**). Notably, at the single-cell level, the 30/60 protocol detected a slightly lower number of lncRNAs ( $388 \pm 29.46$  sd) than in the 37/30 protocol ( $417 \pm 52.61$  sd), albeit higher numbers of protein-coding and processed pseudogenes ( $3,508 \pm 290.04$  sd and  $385 \pm 63.06$  sd vs.  $3,377 \pm 354.13$  sd and  $250 \pm 45.63$  sd) (**Figure 2H**).

Moreover, we characterized the transcriptome coverage along the full 5' to 3' length for both uMRT protocols. For both uMRT protocols, we found similar tendencies on the bulk (**Figure 2I**) as well as on the single-cell level (**Figure 2J**). Notably, the top 10 percentile genes showed the most uniform transcriptome coverage across both protocols on the bulk and single-cell level, with the 30/60 protocol spanning more evenly across the full length of the transcriptome than the 37/30 protocol. Conversely, the upper middle quartile genes showed an increased

coverage at the 3' end. A similar trend was observed for the genes with low expression levels, albeit less pronounced.

Taken together, we demonstrate that uMRT generates a high-quality transcriptome with minimal TSO concatenation (< 3 %) and captures approximately 4,000 genes in the range of other protocols (Mereu *et al.*, 2020). We observed that the 30/60 uMRT protocol represents a better set of conditions, especially on the single-cell level, with lower rates of TSO concatemers, higher percentages of uniquely mapping reads, higher numbers of detected genes, and a slightly more uniform coverage across the transcriptome than the 37/30 protocol.



**Figure 2: uMRT-based scRNA-seq optimization and metrics.**

**(A-B)** Yield (left panel) and Bioanalyzer profiles (right panel) of cDNA libraries prepared from K562 cells using the 30/60 protocol (blue) and the 37/30 protocol (green). Mean is shown as horizontal bar; individual samples are shown as points. Bulk samples represent 1 ng of extracted K562 RNA (n=2). Single cells were sorted by FACS and processed individually (n=9). NC: Negative control.

**(C)** TSO concatemer rates of two uMRT-based protocols. TSO concatemer rates of the 30/60 protocol (left, blue) and the 37/30 protocol (right, green), shown for bulk-level input (1 ng, n=2 for each protocol) and single cells (n=9 for each protocol). Barplots indicate mean percentages; points denote single samples. The percentage of TSO concatemers was calculated by dividing the number of reads containing more than one TSO sequence and its reverse complementary by the total number of reads.

**(D)** Mapping statistics of the 30/60 protocol (left, blue) and the 37/30 protocol (right, green). The percentage of uniquely mapped reads (top-row panels), multiple mapping reads (mid-row panels) and unmapped reads (bottom-row panels). Barplots indicate mean mapping statistics and points denote single samples.

**(E)** Sequencing saturation curve for the 30/60 protocol (blue) and the 37/30 protocol (green) at bulk level. Error bars indicate the standard deviation between replicates (n=2).

**(F)** Sequencing saturation curve as shown in **(E)**, but for single cells. Boxplots show the distribution of the number of detected genes between all samples (n=9).

**(G)** Gene detection sensitivity of the 30/60 protocol (blue) and the 37/30 protocol (green) at bulk level. Depicted are genes with TPM ≠ 0 according to different biotypes (protein coding, lncRNA, processed pseudogene) downsampled to 19 million reads to account for sequencing depth effects. Barplots show the mean between the samples with single sample values shown as points.

**(H)** Gene detection sensitivity as shown in **(G)**, but for single cells.

**(I)** Gene body coverage analysis of the 30/60 protocol (blue) and the 37/30 protocol (green) at bulk level. Shown are the gene body coverage of low expression genes (left), upper middle quartile genes to account for high-coverage genes (center) and the highest expressed genes in the highest percentile (right).

**(J)** Gene body coverage as shown in **(I)**, but for single cells.

## **uMRT shows a high capture efficiency of the mitochondrial transcriptome**

After selecting the optimal uMRT protocol conditions, we compared the performance of the uMRT 30/60 protocol (hereafter referred to as simply uMRT) to Smartseq at the single-cell level. Strikingly, we observed significant differences in read distribution between uMRT and Smartseq for mitochondrial transcriptome coverage. For uMRT, on average, mitochondrial reads accounted for 44.38 % (sd = 4.88) of total captured TPM compared to 3.13 % (sd = 1.67) with Smartseq (**Figure 3A**).

While we know that mitochondrial RNA read proportions vary widely between approximately 10 % in blood and > 50 % in heart tissue (Ludwig *et al.*, 2019), we wanted to examine our findings closely using independent datasets. First, we look back TGIRT-seq dataset (Group II intron based RNA-seq using TGIRT enzyme) from K562 and UHRR (Yao *et al.* 2024), and mitochondrial proportion also reaches >45%. Also, to ensure that this effect does not stem from dead cells, differential cell lysis properties, or other artifacts, we reanalysed two uMRT-based bulk RNA-seq (1 ng as input) datasets on universal human reference RNA (UHRR) generated side by side using uMRT and Superscript II (SSII) (Guo *et al.*, 2024). UHRR contains a mixture of total RNA from ten human cell lines, and reflects transcriptomic

complexity and differences, making it suitable as a comparative benchmark. We also included bulk UHRR data from an unpublished dataset with an updated TSO. For both, we assessed the number of detected genes in relation to the percentage of mitochondrial sequencing reads (**Supplementary Figure 2**). In the dataset of Guo et al., we detected an average of  $19,848 \pm 1,110.16$  sd genes for uMRT and  $17,259 \pm 967.32$  sd genes for SSII at a sequencing depth of 20 million reads (**Supplementary Figure 2A, top panel**). Strikingly, we observed a mean mitochondrial percentage of  $34.41\% \pm 1.11$  sd for uMRT in comparison to only  $1.93\% \pm 0.04$  sd for SSII (**Supplementary Figure 2A, bottom panel**). Similarly, for the two replicates in the UHRR bulk dataset using an updated TSO, we found a mean mitochondrial percentage of  $39.47\% \pm 0.93$  sd (**Supplementary Figure 2B, bottom panel**), while an average of 33,910 genes were captured (**Supplementary Figure 2B, top panel**). These findings confirm that the higher sensitivity for mitochondrial genes is not merely an artifact, but a robust, consistent finding across uMRT protocols. This means that uMRT provides a better reflection of actual gene abundances within a pool of RNA.

Furthermore, out of the 37 annotated human mitochondrial genes (GENCODE 47 human reference set), we examined the 20 most highly expressed mitochondrial genes between uMRT and Smartseq (**Figure 3B**) and calculated the average fold-change of TPM between uMRT and Smartseq (**Figure 3C**). Each mitochondrial gene was detected at higher expression levels in uMRT and changes ranged from 1.5-fold to almost 20-fold, with *MT-RNR1* (18.67-fold), *MT-ND2* (16.37-fold) and *MT-CYB* (15.17-fold) having the highest fold changes and *MT-ND3* (2.5-fold), *MT-TP* (1.73-fold) and *MT-TN* (1.65-fold) the lowest, respectively. Then, we examined the read coverage across the mitochondrial genome. uMRT provided an excellent coverage across the mitochondrial genome (**Figure 3D**). When we compare the coverages offered by uMRT and Smartseq, we found that they follow a very similar pattern.

In order to investigate the coverage patterns in more detail, we zoomed in on the genes *MT-RNR1*, *MT-ND6* and *MT-TN*. *MT-RNR1* shows the highest TPM fold change between uMRT and Smartseq (**Figure 3E**). Interestingly, *MT-ND6*, which has an intermediate fold change (8-times) uMRT shows a second peak in expression around the genomic position 14,500, which is not captured at all by Smartseq (**Figure 3G**). Even for *MT-TN*, which exhibits the lowest fold change between the two RTs, the difference is striking (**Figure 3F**). Overall, uMRT mirrors the overall coverage distribution of Smartseq with significantly higher coverages. Taken together, uMRT shows a higher affinity for capturing mitochondrial genes, making it a powerful tool for elucidating the mitochondrial chromosome.

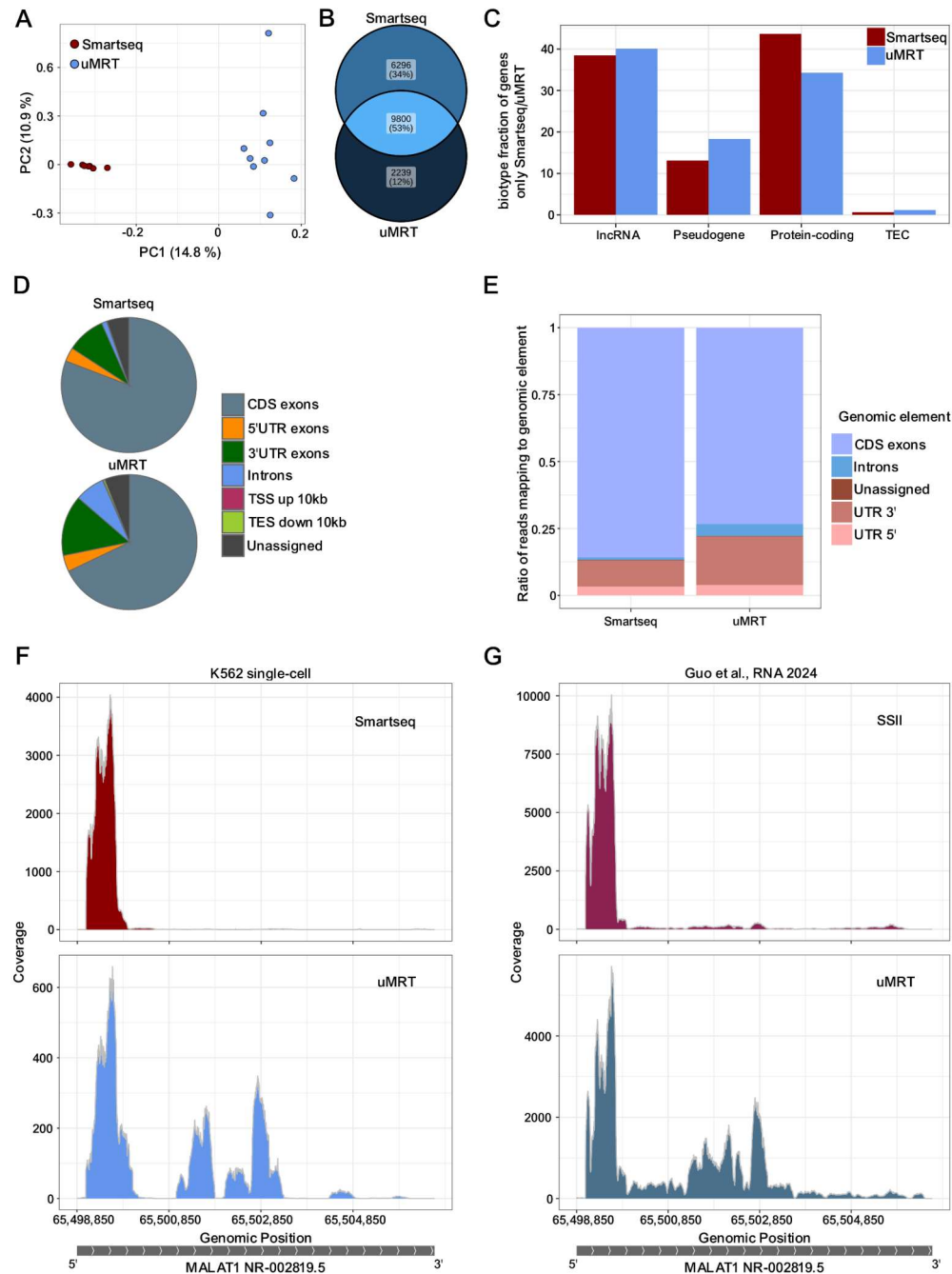


## UMRT captures an enhanced transcriptomic landscape compared to Smart-seq

Next, we sought to examine additional differences in the types of transcripts recognized by the two enzymes. PCA analysis on coding genes only (excluding mitochondrial and residual ribosomal reads) of both enzymes showed a clear separation between uMRT and Smartseq with both enzymes clustering distinctively apart from each other (**Figure 4A**). We observe that uMRT captures different genes from Smartseq, with only 53% of detected genes overlapping between both methods (**Figure 4B**). Interestingly, when examining the distribution of biotypes that are captured by only one of the methods, we observe slight differences. Smartseq captures proportionally more protein-coding sequences (43.7 % of detected genes), whereas the uMRT fraction of protein-coding sequences is approximately 10 % lower (34.3 % of detected genes) (**Figure 4C**). In terms of lncRNA molecules, uMRT and Smartseq capture proportionally similar fractions (38.5 % vs. 40.1 %), although uMRT has a slightly higher tendency to capture processed pseudogenes than Smartseq (18.3 % vs. 13.1 %). Of note, despite their low abundance, uMRT captures twice as many TEC (“to be experimentally confirmed”) genes than Smartseq (1.2 % vs. 0.6 %). This trend was even more pronounced when considering expressed genes (TPM  $\geq$  1) that are captured by only one of the enzymes (**Supplementary Figure 3**). Here, Smartseq captured predominantly protein-coding genes (69.9 % of expressed genes) with uMRT captured proportionally 15 % less (55.2 %). In contrast, uMRT-only expressed genes were further enriched in lncRNA and processed pseudogenes (27.5 % and 10 %) compared to Smartseq (22.7 % and 4.6 %, respectively). Relevant to gene discovery, Smartseq detected virtually no TECs at the expressed level. Also, our data suggests the range of genetic elements profiled by uMRT is more evenly distributed, whereas Smart-seq is highly skewed towards protein-coding segments of genes. This holds true for the totality of genes captured by each method, as well as the overlapping fraction of genes (**Figure 4D, E**). Similar to the trends observed previously, uMRT captures a proportionally smaller fraction of coding sequences (67.91 %) relative to Smartseq (80.8 %). by contrast, uMRT captures on proportionally higher fractions of 3' UTRs and introns than Smart-seq (14.53 % vs. 9.24 % and 7.1 % vs. 1.24 %, respectively). This effect is also pronounced in the genes that overlap between both enzymes. While those genes are captured by both enzymes, Smartseq captures predominantly the coding regions of the genes, whereas uMRT identifies and amplifies all regions of these genes and their isoforms (**Figure 4E**). This is well illustrated by the lncRNA MALAT1. Even though this gene is ubiquitously and abundantly expressed in cells and tissues, the two enzymes exhibit markedly different capture patterns. While the overall coverage is lower for uMRT than for Smartseq, uMRT elucidates downstream exons located towards the 3' end that remain entirely invisible for Smartseq (**Figure 4F**). Comparable results were also observed on the bulk level when comparing 1 ng

of UHRR between uMRT and SSII from the reanalysed data of Guo et al. (Guo et al., 2024). Additionally, the same trend was evident with uMRT with inputs as low as 2 pg of UHRR (**Supplementary Figure 4**), reinforcing the capacity of uMRT to elucidate new regions of the genome.

In summary, uMRT captures an enhanced transcriptomic landscape compared to state-of-the-art reverse transcriptases and can illuminate regions of the genome that have previously remained inaccessible.



**Figure 4: MMLV and uMRT capture different genomic structures.**

**(A)** Principle component analysis (PCA) of genes driving variability between Smartseq (red) and uMRT (blue). PCA was calculated on the 2000 most variable genes, using 10 principle components. Mitochondrial and ribosomal genes have been filtered out.

**(B)** Intersection of expressed genes averaged across all cells captured by Smartseq only (top, blue) and uMRT 30/60 protocol only (bottom, darkblue), as well as both methods (center, lightblue). Genes are considered detected at TPM > 0.

**(C)** Distribution of biotypes in percent between genes detected only in Smartseq (red) and only in uMRT (blue). Biotypes are shown above 0.1 % of the total expression.

**(D)** Pie charts depicting the proportional distribution of mapped sequencing reads across genomic features captured by Smartseq (top) and uMRT (bottom), averaged across single cells. Pie slices are coloured according to genomic features.

**(E)** Ratio of genomic elements in overlapping genes between uMRT and Smartseq. The ratio of genomic elements of overlapping genes was calculated by extracting reads of each genomic element and comparing it to the amount of total reads. Genomic elements, as extracted from RSeQC, are indicated in different colours.

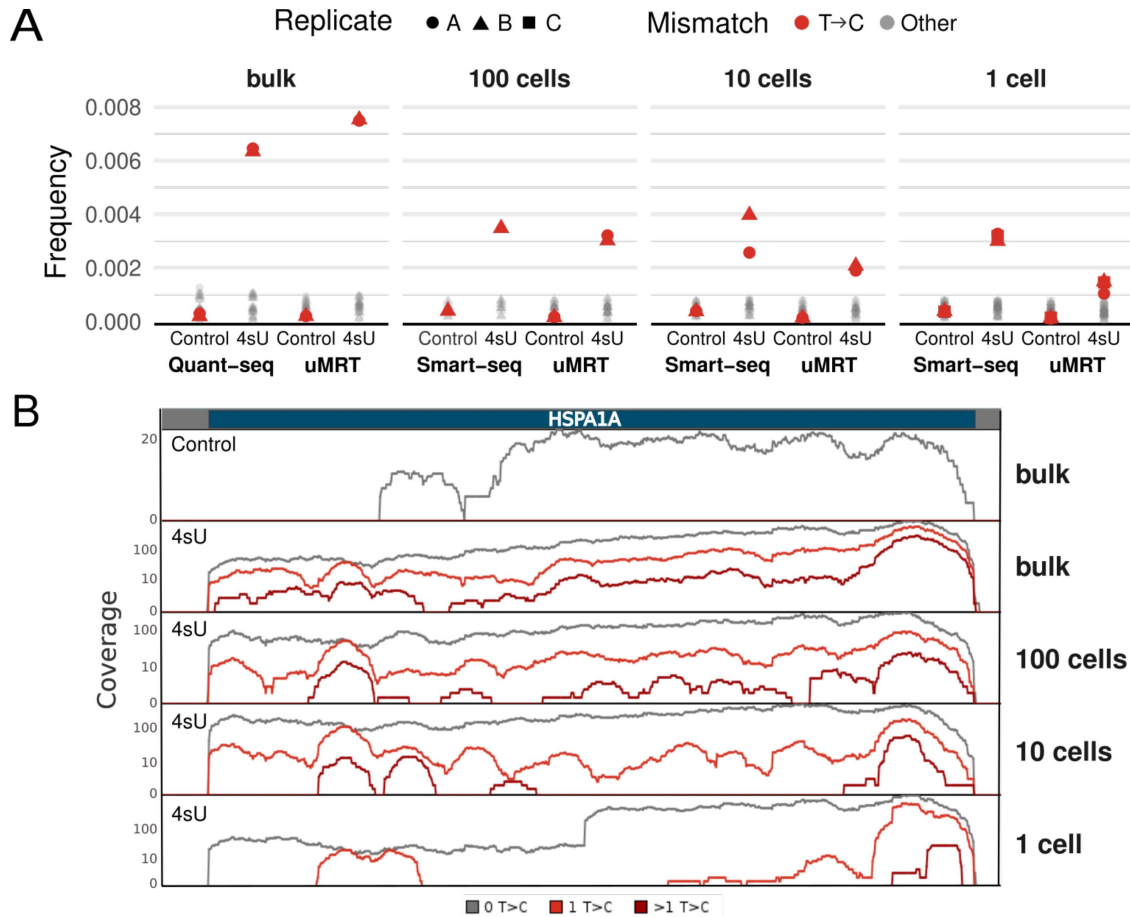
**(F)** Coverage tracks of the *MALAT1* gene of K562 single cells. Coverage tracks depict the averaged coverage of Smartseq reads (top panel, red) and of uMRT reads (bottom panel, blue) averaged across single cells. The gene model (grey box) depicts the RefSeq *MALAT1* NR-002819.5 transcript.

**(G)** Coverage tracks of the *MALAT1* gene from the reanalyzed bulk UHRR dataset of Guo et al., 2024. Coverage tracks depict the average coverage of SSII reads (n=2, top panel, magenta) and of uMRT (n=2, bottom panel, petrol blue). The gene model (grey box) depicts the RefSeq *MALAT1* NR-002819.5 transcript, as shown in **(F)**.

## uMRT Enables Detection of Newly Synthesized RNA

Given the sensitivity of our methods, we wanted to evaluate the ability of uMRT to detect newly synthesized RNA using RNA metabolic labeling and nucleotide conversion-based approaches (Erhard *et al.*, 2019). We applied metabolic labeling with 4-thiouridine (4sU) on HeLa cells at a concentration of 300  $\mu$ M for 2 hours, and compared mismatch profiles across three sequencing technologies: Quant-seq, Smart-seq, and uMRT. Using matched labeled and unlabeled samples across a range of input sizes (bulk, mini-bulk, and single cell; see Methods), we were able to quantify the number of mismatches (Jürges, Dölken and Erhard, 2018). Specifically, we observed that T→C mismatches, indicative of 4sU incorporation, were enriched in labeled samples across all technologies, while control samples showed low background levels (**Figure 5A; Supplementary Figure 5**). This demonstrates that uMRT retains a specific T→C signal down to the single-cell level that can be used to differentiate newly synthesized from pre-existing RNA upon 4sU labeling and nucleotide conversion (**Figure 5A**). A focus on the genome browser of the *HSPA1A* locus further illustrates this capability (**Figure 5B**): T→C mismatches are clearly detectable in uMRT data, even at the single-cell level. Importantly, both single and multiple T→C events can be resolved. Interestingly, most other mismatch frequencies are lower for uMRT than for Smart-seq, underscoring the high-fidelity reverse transcription of uMRT (**Supplementary Figure 5**). This is particularly the case for T→C mismatches in the unlabeled control samples (0.03788 % for Smart-seq vs. 0.01419 % for uMRT), which is highly beneficial in preventing the overestimation of labeled RNA (Jürges, Dölken and Erhard, 2018). Taken together, these results demonstrate that

uMRT can detect 4sU-induced T→C mismatches across a range of input sizes, indicating that uMRT is suitable for metabolic RNA labeling and holds promise for studying transcriptional dynamics.



**Figure 5. uMRT Detects 4sU-Induced T→C mismatches across bulk, mini-bulk and single-cell inputs.**

**(A)** Scatterplot of global mismatch frequencies per sample (HeLa cells) in sense reads across different sequencing technologies (Quant-seq, Smart-seq, uMRT). T→C mismatches are shown in red, all other mismatch types are shown in grey. Shapes indicate replicates.

**(B)** Genome browser view of HSPA1A coverage, highlighting T→C mismatches. The top panel shows control (unlabeled) bulk data. All other panels show 4sU-labeled samples across decreasing input sizes (bulk (1ng) to 1 cell). Total coverage is shown in black, coverage of reads containing exactly one T→C mismatch is shown in red, and coverage with more than one T→C mismatch is shown in dark red.

## Discussion

The ability to copy RNA molecules into complementary DNA molecules using RTs has revolutionized the ability to profile RNA molecules (Stark, Grzelak and Hadfield, 2019). The

success of MMLV-based RNA-seq workflows lies on their ability to perform template-switching, enabling versatile use across various protocol formats (e.g. plate-based, droplet-based), and facilitating easy-to-implement workflows with minimal experimental steps (Hahaut *et al.*, 2022). This capability also enables picogram level of sensitivity, paving the way for performing scRNA-seq (Picelli *et al.*, 2013). By systematically utilizing different experimental approaches through iterative experiments and 20 years of protocol optimization, some workflows now achieve exceptional transcript detection rates with over 10,000 transcripts per cell (Hagemann-Jensen *et al.*, 2020; Hagemann-Jensen, Ziegenhain and Sandberg, 2022; Hahaut *et al.*, 2022). Yet, the most significant developments have continuously focused on bringing scRNA-seq to a larger scale in order to reach organism-level profiling of mRNA, which comes at the expense of capturing the complete census of the single-cell transcriptomes (Quake, 2021).

MMLV RT enzymes have inherent processivity limitations and are unable to unfold complex RNA molecules, which introduces bias in RNA-seq (Verwilt, Mestdagh and Vandesompele, 2023). Nonetheless, these artifacts have often been overlooked because no RT alternatives were available. Since the introduction of reverse transcriptases derived from group II intron retroelements as biotechnology tools (Mohr *et al.*, 2013; Zhao, Liu and Pyle, 2018), and especially the proof that these enzymes can perform RNA-seq (Nottingham *et al.*, 2016; Guo *et al.*, 2024), it is possible to envision developing a de-novo workflow for scRNA-seq. On the bulk level (meaning using ng total RNA quantities), group II intron-based RNA-seq has demonstrated the capability to capture an unforeseen repertoire of coding and non-coding RNA molecules (Boivin *et al.*, 2018; Guo *et al.*, 2024). Yet, major roadblocks had to be lifted in order to reach single-cell sensitivity, including the ability to handle picogram levels and achieve efficient amplification of cDNA. Current group II intron-based RNA-seq protocols have only been demonstrated to handle nanogram levels of purified RNA (Boivin *et al.*, 2018; Guo *et al.*, 2024; Unlu *et al.*, 2024). Here, for the first time to our knowledge, we have harnessed a group II intron RT, UltraMarathonRT (Guo *et al.*, 2024), to reach picogram-level sensitivity, gaining a three to four orders of magnitude level of sensitivity compared to current protocols (**Figure 1**). In addition, we systematically improved every step of the workflow (**Figure 1A**) to maximize the ability of uMRT to perform template-switching and amplify picogram-level RNA (**Figure 1B-F**). We then successfully performed transcriptome-wide RNA-seq (**Figure 2**), reaching thousands of genes per single cell and overall performance that is comparable to standards in the field (Mereu *et al.*, 2020). Importantly, the template-switching protocol has not led to the accumulation of TSO concatemers (**Figure 2**), unlike other scRNA-seq protocols (Hagemann-Jensen *et al.*, 2020; Kapteyn *et al.*, 2010).

Having overcome the technical barriers involved in implementing uMRT scRNAseq workflow, we can now see the rich landscape of RNAs that can be visualized by a high performance RT at the single cell level. Perhaps not surprisingly, we observe that the transcriptome space captured by uMRT varies substantially compared to MMLV RTs (**Figure 3, 4**). The finding of most general relevance is that uMRT reveals a repertoire of coding genes that were previously undetected (**Figure 4A-B**), with more than 2,000 genes that have been uniquely identified by this enzyme. In this protocol, a homogenous 5' to 3' coverage profile is observed for abundant RNA molecules (**Figure 2I, J**), while a 3' bias is observed for the least abundant RNAs (**Figure 2I, J**). Nonetheless, uMRT maintains a greater ability to identify the diversity of genetic elements within genes compared to MMLV-based enzymes, which is similar to our findings in bulk (Guo *et al.*, 2024). This ability to dissect gene composition is reflected in the data on the non-coding RNA *MALAT1*, which is an 8-kilobase-long RNA molecule that is rich in stable RNA secondary structures (Monroy-Eklund *et al.*, 2023). Here we report a far more complex expression profile for *MALAT1* in uMRT-based data relative to MMLV-based data (**Figure 4F-G**).

An unexpected finding is that the new uMRT protocol drastically improves the profiling of mitochondrial poly(A) RNA, increasing sensitivity by a factor of 10 (**Figure 3**). The mitochondrial genome encodes 37 genes, giving rise to abundant polyadenylated mtRNA molecules (Jedynak-Slyvka, Jabczynska and Szczesny, 2021). These can represent up to 50% of the total transcripts in a lysate mixture (Ludwig *et al.*, 2019) and account for the most abundant transcripts (Boivin *et al.*, 2018). Despite this abundance, mtRNA molecules are exceptionally challenging to profile and their characterization has required complicated biochemical steps (Miller *et al.*, 2022; Moran *et al.*, 2024). This has caused significant technological issues, as mtRNA profiling is critical for successful lineage tracing. Given its ability to faithfully report mtRNA populations and isoforms, uMRT has the potential to pave the way for a new generation of assays that overcome the conventional burden of combining scATAC-seq and scRNA-seq (Ludwig *et al.*, 2019; Miller *et al.*, 2022). However, since our current workflow lacks unique molecular identifiers (UMIs), it will be important for us to determine whether there is a bias in amplification that leads to an overrepresentation of mtRNA.

Finally, we aimed to demonstrate that uMRT is compatible with time-resolved scRNA-seq using metabolic RNA labeling (**Figure 5**) (Erhard *et al.*, 2022). Conceptually, we used 4-thiouridine (4sU) as a nucleotide analog that can be incorporated into newly synthesized RNA molecules, leading to T-C mismatches in the reads. We compared the Quant-seq workflow that was originally used to perform SLAM-seq (Herzog *et al.*, 2017), Smart-seq using Takara kits

at the single-cell level (Erhard *et al.*, 2019), and uMRT-based workflows. We established that 4sU incorporation can be effectively detected with uMRT down to the single-cell level. A significant challenge in data analysis for such experiments is distinguishing bona-fide, 4sU induced T→C mismatches from mismatches induced by other means, such as low-fidelity RT. Here, we found that T→C mismatch frequencies in control samples were significantly lower for the uMRT than for the Smart-seq libraries. Altogether, it demonstrates that uMRT is not only amenable to unveil the transcriptome at the single-cell level but also its dynamics.

## Methods

### Total RNA extracts and synthetic RNA

Commercially available RNA mixes were derived from mouse brain (Takara Bio, #636601) or mixed human cell lines Universal Human Reference RNA (UHRR, ThermoFisher, #QSO639). Synthetic RNA from Spike-In RNA Variant Set 4 (SIRV-Set 4, Lexogen, #141.01) was used for targeted amplification with uMRT. All RNA samples were stored at stock concentrations in aliquots at -80°C for one-time use. Dilutions were prepared using nuclease-free water (Ambion) containing 0.4 units/μL recombinant RNase inhibitor (Takara Bio). RNA sample quality was monitored by the Bioanalyzer RNA Pico assay (Agilent).

### Cell lines

ENCODE tier-one cell line K-562 (ATCC CCL-243) and tier-two cell line HeLa S3 (ATCC CCL-2) were purchased from ATCC. K562 cells were cultured in RPMI 1640 medium (Gibco) and HeLa S3 cells were cultured in DMEM (Gibco). Both types of culture media were supplemented with 10% v/v FBS (Fetal Bovine Serum, Gibco), 1% v/v penicillin-streptomycin (100 units/mL penicillin and 100 μg/mL streptomycin, Gibco), and 2mM L-glutamine (Gibco). Cells were passaged every second day and cultured at 37°C with 5% CO<sub>2</sub> saturation. Mycoplasma absence from cell culture was regularly ensured using the Mycoplasma check service of Eurofins Genomics.

### Low-input target amplification using uMRT

Targeted capture and amplification were performed on the SIRV4001 transcript from SIRV-Set 4 (SIRV-Set 4, Lexogen, #141.01). For the carrier RNA assays, 10 pg SIRV-Set 4 were mixed with a 10-fold dilution series of total mouse brain RNA (1 μg to 10 pg, Takara Bio) in a total

volume of 1  $\mu\text{L}$ . In absence of total mouse brain RNA, 10-fold dilution series of SIRV-Set 4 were prepared from 10 pg to 1 fg in 1  $\mu\text{L}$  volumes. Then, 2  $\mu\text{L}$  of 1  $\mu\text{M}$  oligodT3oVN primer (5'-TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN-3', IDT), 1  $\mu\text{L}$  of dNTP mix (10 mM each, NEB) and 2  $\mu\text{L}$  nuclease-free water (Ambion) were added to a volume of 6  $\mu\text{L}$ . The reaction was incubated at 95°C for 30 sec, then snap-cooled on ice to anneal the primer to the template. For reverse transcription, 10  $\mu\text{L}$  of RT buffer (100 mM Tris-HCl, pH 8.3 at 25°C (Invitrogen), 400 mM KCl (Ambion), 4 mM MgCl<sub>2</sub> (Ambion), 10 mM DTT (Sigma), 40 % glycerol (Promega)), 1  $\mu\text{L}$  of uMRT (20 units/ $\mu\text{L}$ , RNAConnect), 0.5  $\mu\text{L}$  of RNaseOUT (ThermoFisher), 1  $\mu\text{L}$  Boost (if added, otherwise nuclease-free water, RNAConnect) and 1.5  $\mu\text{L}$  nuclease-free water (Ambion) were added to a total volume of 20  $\mu\text{L}$ . The reaction was incubated at 42°C for 60 min and inactivated at 95°C for 1 min. Next, 1  $\mu\text{L}$  of 8  $\mu\text{M}$  PCR\_fwd01 primer (5'-GAAGTTAACAAGTAGAATCATTATCCGGC-3', IDT), 1  $\mu\text{L}$  of 8  $\mu\text{M}$  PCR\_rev01primer (5'-TACGCTTACTCCAATACGACGCTC-3', IDT), 10  $\mu\text{L}$  LongAmp Taq 2X Master Mix (NEB), 7  $\mu\text{L}$  nuclease-free water were added. The cDNA was amplified using the following program: initial denaturation at 94°C for 3 min, then 94°C for 30 sec, 56°C for 30 sec, 65°C for 5 min for 36 cycles; final extension at 65°C for 10 min. The PCR products were cleaned up with magnetic AMPure XP beads (Beckman Coulter) in a 0.7x ratio. The cDNA was quantified with Qubit 1X dsDNA High Sensitivity (ThermoFisher), and profiled on the Bioanalyzer using the High Sensitivity DNA Kit (Agilent).

### **Single-cell sorting on plates**

K562 cell suspensions were collected, washed twice in 1x PBS (Gibco) by centrifuging at 250g for 5 min, and resuspended in 1x PBS (Gibco) to reach a concentration of  $1.10^6$  cells/100  $\mu\text{L}$ . Live-dead cell staining assay was performed with 5  $\mu\text{L}$  of Propidium Iodide Ready Flow Reagent (ThermoFisher) added to 100  $\mu\text{L}$  cell suspension and incubated in the dark for 2 min on ice. The stained cell suspension was diluted three times in 1x PBS (Gibco) and passed through a 70- $\mu\text{m}$  cell strainer (pluriSelect) before loading on the FACSaria Fusion Flow Cytometer (BD Biosciences). After setting up gating that ensures that dead cells and doublets are excluded, single cells were sorted through a 100  $\mu\text{m}$  nozzle into cooled, 48-well no-skirt PCR plates (Brand) prefilled with lysis buffer (see below). Immediately after sorting, each plate was spun down at maximum speed (500 g), transferred on dry ice, and frozen at -80°C. Plates containing single cells were thawed on ice for 3 min and spun down before use.

## **uMRT-based RNA-seq and scRNA-seq**

***Cell lysis buffer and annealing.*** The transcriptome-wide capture assays used 1.5  $\mu$ L lysis buffer consisting of 1x Takara Lysis Buffer (Takara Bio), 0.22 units/ $\mu$ L of recombinant RNase inhibitor (Takara Bio) and 2.4 M betaine (Thermo Fisher). For assays using extracted total mouse brain RNA, the RNA was diluted to 1 ng (bulk) or 30 pg (single cell-level) and added into wells of 0.2 mL PCR tubes containing 1.5  $\mu$ L of lysis buffer. Plates for single cell-sorting (K562 and HeLa S3 cells) were prefilled with 1.5  $\mu$ L lysis buffer before sorting. For annealing, 1.25  $\mu$ M uMRT\_dT18 primer (5'-CCCTCTCTCTCTTTTCCTCTCTCTTTTTTTTTTTTTTTTTTT-3', IDT) and 1.25 mM dNTPs (NEB) were added in a final volume of 2.4  $\mu$ L and incubated at 95°C for 15 seconds, followed by snap-cooling on ice.

***Reverse transcription.*** Next, the primer-annealed RNA was reverse transcribed in a final reaction volume of 6  $\mu$ L containing 50 mM Tris-HCl, pH 8.3 at 25°C (Invitrogen), 200 mM KCl (Ambion), 4 mM MgCl<sub>2</sub> (Ambion), 5 mM DTT (Sigma), 20 % glycerol (Promega), 0.5 U/ $\mu$ L Recombinant RNase Inhibitor (Takara Bio), 1x Boost (RNACONNECT) and 0.33 units/ $\mu$ L uMRT (RNACONNECT). The reaction was incubated at 42°C for 30 min.

***Template-switching.*** Once reverse transcription was completed, 25 mM Tris-HCl pH 8.3 at 25°C (Invitrogen), 2 mM MgCl<sub>2</sub> (Ambion), 2.5 mM DTT (Sigma), 10% PEG4000 (Sigma), 1  $\mu$ M TSO (5'-CCCTCTCTCTCTTTTCCTCTCTCTTTT-3', RNACONNECT), 1 mM dATP (Carl Roth) and 0.67 units/ $\mu$ L uMRT (RNACONNECT) were added to a final reaction volume of 12  $\mu$ L and incubated at 42°C, 34°C or 30°C for 60 min for template-switching, followed by inactivation at 80°C for 5 min.

***PCR amplification.*** cDNA was amplified in a volume of 50  $\mu$ L containing 1x KAPA GC enhanced buffer (Roche), 0.02 U/ $\mu$ L KAPA HiFi DNA polymerase (Roche), 0.3 mM dNTPs (Roche) and 2  $\mu$ M AmpPCR primer (5'-CCCTCTCTCTCTTTTCCTCTCTC-3', IDT). After initial denaturation at 98°C for 2 min, the program was cycled 25-times at 98°C, 15 s, 62°C, 30 s and 72°C, 6 min, for a final extension at 72°C for 5 min. During optimization, annealing temperatures between 64°C-58°C were tested in 2°C steps as indicated in **Supplementary Figure 1A**. After PCR, the reaction was cleaned-up with AMPure XP beads (Beckman Coulter) in a 0.5x ratio and eluted in elution buffer (Qiagen). The PCR product concentrations were quantified with Qubit 1X dsDNA High Sensitivity (ThermoFisher), and length distribution was profiled on the Bioanalyzer using the High Sensitivity DNA Kit (Agilent).

## **SMART-seq based RNA-seq and scRNA-seq**

SMART-seq libraries were prepared with Takara Bio's v4 Ultra Low Input RNA Kit according to the manufacturer's instructions with a quarter of the recommended reagent volumes. In brief, cells were sorted into plates prefilled with 2.6  $\mu\text{L}$  lysis buffer consisting of 1x Lysis buffer (Takara Bio) and nuclease-free water (Ambion) and stored at  $-80^{\circ}\text{C}$ . Upon use, the plates were thawed on ice, and 1.4  $\mu\text{L}$  of first strand primer mix consisting of 0.2  $\mu\text{L}$  RNase inhibitor, 0.5  $\mu\text{L}$  CDS primer, and 0.7  $\mu\text{L}$  nuclease-free water was added to each well. Cell lysis and annealing were performed by incubating for 3 min at  $72^{\circ}\text{C}$  in a pre-heated thermocycler, followed by cooling down at  $4^{\circ}\text{C}$  for 3 min. Reverse transcription readily followed by adding 1  $\mu\text{L}$  of 5X Ultra-low first strand buffer, 0.25  $\mu\text{L}$  of SMART-seq v4 Oligo, 0.125  $\mu\text{L}$  RNase inhibitor, 0.5  $\mu\text{L}$  SMARTScribe Reverse Transcriptase and incubating at  $42^{\circ}\text{C}$  for 90 min, then at  $70^{\circ}\text{C}$  for 10 min. Second strand synthesis and amplification was performed in 6.25  $\mu\text{L}$  2X SeqAmp PCR Buffer, 0.25  $\mu\text{L}$  PCR Primer II A, 0.25  $\mu\text{L}$  SeqAmp DNA Polymerase and 0.75  $\mu\text{L}$  nuclease free-water at  $95^{\circ}\text{C}$  for 1 min, 21 cycles at  $98^{\circ}\text{C}$  for 10 s,  $65^{\circ}\text{C}$  for 30 s,  $68^{\circ}\text{C}$  for 3 min and  $72^{\circ}\text{C}$  for 10 min for final extension. cDNA libraries were cleaned-up by adding 12.5  $\mu\text{L}$  (1:1 beads ratio) AMPure XP beads (Beckman Coulter) and 0.25  $\mu\text{L}$  1x Lysis buffer (Takara Bio) and incubating for 8 min at room temperature. Next, the beads were washed twice for 30 s in 100  $\mu\text{L}$  80% ethanol, before air-drying and eluting in 17  $\mu\text{L}$  elution buffer (Takara Bio). cDNA was quantified using Qubit 1X dsDNA High Sensitivity (ThermoFisher), and profiled with the Bioanalyzer High Sensitivity DNA assay (Agilent).

## **Library preparation and sequencing**

Sequencing libraries were prepared using the Nextera XT DNA Library Prep Kit (Illumina). In brief, 0.5 ng cDNA was tagmented using 2  $\mu\text{L}$  Tagmentation DNA Buffer and 1  $\mu\text{L}$  Amplicon Tagment Mix in a thermocycler at  $55^{\circ}\text{C}$  for 15 min, with a hold step at  $10^{\circ}\text{C}$ . Immediately after the samples reach  $10^{\circ}\text{C}$ , 1  $\mu\text{L}$  neutralizing buffer was added and incubated for 5 min at room temperature to strip off the Tn5 enzyme. The fragments were enriched with 3  $\mu\text{L}$  Nextera PCR master mix, 1  $\mu\text{L}$  Index 1 primer and 1  $\mu\text{L}$  Index 2 primer in total reaction volume of 10  $\mu\text{L}$ . The PCR program comprised denaturation at  $72^{\circ}\text{C}$  for 3 min and at  $95^{\circ}\text{C}$  for 30 s, followed by 12 cycles of denaturation at  $95^{\circ}\text{C}$  for 10 s, annealing at  $55^{\circ}\text{C}$  for 30 s, extension at  $72^{\circ}\text{C}$  for 30 s. Finally, a final extension was performed at  $72^{\circ}\text{C}$  for 5 min. The libraries were cleaned up with AMPure XP beads in a 0.6x ratio (Beckman Coulter), resuspended in 13  $\mu\text{L}$  Resuspension buffer (Illumina). Library concentration was measured using the Qubit 1X dsDNA High Sensitivity or Quant-iT PicoGreen dsDNA Assay (ThermoFisher), and library size was visualized using the Bioanalyzer High Sensitivity DNA assay (Agilent). Finally, libraries were

sequenced at the 100-bp single end or 150-bp paired end on an Illumina NextSeq2000 or NovaseqX for 11 million reads (single cells) or 20 million reads (bulk).

### scSLAM-seq

***4-thiouridine labeling.*** One day prior to the experiment,  $5 \times 10^5$  HeLa S3 cells were seeded in a T25 flask to reach confluency the next day. On the day of the experiment, 30  $\mu\text{L}$  of 50 mM 4-thiouridine (Biosynth) was added to 5 mL DMEM medium (Gibco) and incubated for 2 hours. Then, cells were harvested by detaching with 0.5 mL Trypsin-EDTA (0.25%, Gibco) for 5 min. Trypsin treatment was stopped with 4.5 mL DMEM medium and cells were washed twice in 1x PBS (Gibco) by centrifuging at  $250 \times g$  for 5 min. Cell sorting (single cells, 10 cells, and 100 cells) on plates was performed as described above. For bulk SLAM-seq,  $5 \times 10^4$  cells were collected per sample and centrifuged at  $4,500 \times g$  for 3 min. The pellet was then resuspended in 100  $\mu\text{L}$  RNA lysis buffer (Zymo Research) and stored at  $-80^\circ\text{C}$  for further processing.

***Alkylation of 4-thiouridine.*** Wells containing single cells in 1.5  $\mu\text{L}$  lysis buffer for uMRT or 2.6  $\mu\text{L}$  lysis buffer for Smart-seq were filled to 4  $\mu\text{L}$  with nuclease-free water (Ambion). 0.4  $\mu\text{L}$  of 10x PBS pH 7.4 (Gibco) were added to all wells. Then, 4.4  $\mu\text{L}$  of 20 mM Pierce™ Iodoacetamide (ThermoFisher) resuspended in DMSO (Sigma) was added to all wells and briefly centrifuged. Alkylation took place during an incubation of 5 min at  $50^\circ\text{C}$  in a thermocycler. The reaction was quenched with 1.3  $\mu\text{L}$  of 0.1 M DTT (ThermoFisher) for 5 min at room temperature.

***Alkylation of 4-thiouridine for bulk samples.*** RNA was extracted using the Quick-RNA Microprep Kit (Zymo Research) according to the manufacturer's instructions. In brief, 100  $\mu\text{L}$  of absolute ethanol was added to 100  $\mu\text{L}$  RNA sample in RNA lysis buffer. After mixing, the sample was spun through a Zymo-Spin™ IC Column at  $16,000 \times g$  for 30 sec. Next, each sample was washed with 400  $\mu\text{L}$  RNA Wash Buffer, centrifuged and treated with 40  $\mu\text{L}$  DNase I mixture on the column for 15 min. After DNase I treatment, the samples were washed with 400  $\mu\text{L}$  RNA Prep Buffer, 700  $\mu\text{L}$  RNA Wash Buffer and 400  $\mu\text{L}$  RNA Wash Buffer with centrifugation at  $16,000 \times g$  between each step. Finally, the RNA was eluted in 15  $\mu\text{L}$  nuclease-free water (Ambion). RNA quantity was measured using the Qubit™ RNA Broad Range (BR) kit and RNA quality was visualized using the Bioanalyzer RNA 6000 Pico assay (Agilent). After validating the RNA samples, 5  $\mu\text{L}$  of extracted RNA was added to 2.2  $\mu\text{L}$  nuclease-free water (Ambion) on ice. Then, 1.8  $\mu\text{L}$  of 10x PBS pH 7.4 (Gibco) was added to each sample, followed by 9  $\mu\text{L}$  of 20 mM Pierce™ Iodoacetamide (ThermoFisher) resuspended in DMSO (Sigma). The samples were mixed and incubated at  $50^\circ\text{C}$  for 15 min in a thermocycler. The alkylation was quenched adding 3.2  $\mu\text{L}$  of 0.1 M DTT (ThermoFisher) and incubating for 5 min at room temperature.

Clean-up step for sorted cell samples. 10.6  $\mu\text{L}$  of RNA XP beads (1:1 ratio, Beckman Coulter) were added to the reaction, mixed well and incubated for 10 min at room temperature. Then, the samples were placed on a magnet for 4-5 min until the supernatant became clear. The supernatant was discarded, and the beads were washed twice for 30 s with 50  $\mu\text{L}$  of 80% ethanol. After the beads were air-dried for a maximum 2 min, the RNA was resuspended and incubated for 3 min at room temperature in elution buffer consisting of the components for annealing depending on the protocol (uMRT or SMART-seq). For elution, the samples were centrifuged at 500 x g for 5 min until the beads were pelleted at the bottom of the wells. From that point on, samples were centrifuged at 500 x g for 5 min before every incubation step to avoid bead interference with downstream reactions.

Clean-up step for bulk samples. Alkylated samples were cleaned using the Quick-RNA Microprep Kit (Zymo Research) according to the manufacturer's instructions. In brief, 64  $\mu\text{L}$  of RNA lysis buffer was added to the alkylation reactions. Then, the lysate was mixed with 85  $\mu\text{L}$  of absolute ethanol and loaded on a Zymo-Spin IC Column. The remaining steps were performed as described above. The samples were eluted in 11  $\mu\text{L}$  nuclease-free water (Ambion) and the Qubit RNA Broad Range (BR) kit was used to measure RNA quantity, while the RNA quality was visualized using the Bioanalyzer RNA 6000 Pico assay (Agilent).

uMRT-based RNA-seq. The RNA beads were resuspended and incubated for 3 min at room temperature in 1.68  $\mu\text{L}$  elution buffer consisting of 0.78  $\mu\text{L}$  nuclease-free water (Ambion), 0.6  $\mu\text{L}$  of 5  $\mu\text{M}$  uMRT\_dT18 primer (5'-CCCTCTCTCTCTTTCTCTCTTTTTTTTTTTTTTTTTTT-3', IDT), 0.3  $\mu\text{L}$  dNTP mix (10 mM each, NEB). After centrifuging at 16,000g for 5 min, 0.72  $\mu\text{L}$  of 5 M betaine (ThermoFisher) was added. The remaining steps were performed as described above. The samples were sequenced for 5 million reads (single cells, 10 cells, and 100 cells) and 10 million reads (bulk).

SMART-seq. The RNA beads were resuspended and incubated for 3 min at room temperature in 4  $\mu\text{L}$  elution buffer consisting of 0.2  $\mu\text{L}$  RNase inhibitor, 0.5  $\mu\text{L}$  CDS primer, and 3.3  $\mu\text{L}$  nuclease-water. After centrifuging at 16,000 x g for 5 min, all consecutive steps were performed as described above. Single cells, 10 cells, and 100 cells samples were sequenced at 5 million reads, whereas bulk samples were sequenced at 10 million reads.

Quant-seq for bulk samples. Quant-seq libraries were prepared with Lexogen's QuantSeq 3' mRNA-Seq V2 Library Prep Kit FWD according to the manufacturer's instructions. Briefly, 5  $\mu\text{L}$  FS1 was added to 5  $\mu\text{L}$  sample containing 10 ng of extracted and alkylated RNA, and

denatured at 85°C for 3min. Then, the samples were kept at 42°C, while a master mix of 9.5 µL FS2 and 0.5 µL E1 per reaction was prepared. This master mix was pre-warmed at 42°C for 2-3 min before 10 µL of it was added to the samples directly on the thermocycler for first-strand synthesis at 42°C for 15min. Next, RNA was removed by adding 5 µL RS and incubating at 95°C for 10 min, followed by cooling down to 25°C. The first part of second-strand synthesis was performed by adding 10 µL USS to each sample and incubating at 98°C for 1 min, slow-cooling to 25°C at a ramp speed of 0.5°C per sec and incubating at 25°C for 30 min. The second part of second-strand synthesis followed by adding 5 µL of a master mix consisting of 4 µL SS2 and 1 µL E2 and incubating at 25°C for 15 min. After this, the cDNA was purified as follows. 16 µL PB was added, mixed well and incubated for 5 min at room temperature. Then, the samples were placed on a magnet for 2-5 min until the supernatant became clear. The supernatant was discarded and the beads were resuspended in 40 µL EB, followed by incubation for 2 min at room temperature. Then, 56 µL PS was added, mixed well and incubated for 5 min at room temperature. Once more, the samples were placed on a magnet for 2-5 min until the supernatant became clear. The supernatant was discarded again, and the beads were washed twice for 30 sec with 120 µL of 80% ethanol. After being air-dried for a maximum of 5-10 min, the beads were resuspended in 20 µL EB and incubated for 2 min at room temperature. 17 µL of eluted sample was transferred to a fresh PCR plate. For indexing, 8 µL of a mix containing 7 µL PM and 1 µL PE as well as 5 µL of a UDI primer pair was added to each eluted library. Amplification was performed at 98°C for 30 sec, 20 cycles at 98°C for 10 sec, 65°C for 20 sec, 72°C for 30 sec and final extension at 72°C for 1 min. Indexed libraries were cleaned-up in a similar way as before. 30 µL PS was mixed well with the samples and incubated for 5 min at room temperature. The samples were placed on a magnet for 2-5 min until the supernatant became clear. The supernatant was discarded and the beads were resuspended in 30 µL EB, followed by incubation for 2 min at room temperature. Then, 30 µL PS was added, mixed well and incubated for 5 min at room temperature. Once more, the samples were placed on a magnet for 2-5 min until the supernatant became clear. The supernatant was discarded again, and the beads were washed twice for 30 sec with 120 µL of 80% ethanol. After being air-dried for a maximum of 5-10 min, the beads were resuspended in 20 µL EB and incubated for 2 min at room temperature. 17 µL of eluted sample was transferred to a fresh PCR plate. Library concentration was measured using the Qubit 1X dsDNA High Sensitivity or Quant-iT PicoGreen dsDNA Assay (ThermoFisher), and library size was visualized using the Bioanalyzer High Sensitivity DNA assay (Agilent). Finally, bulk libraries were sequenced at the 100-bp single end on an Illumina NextSeq2000 for 10 million reads.

## Data analysis of RNA-seq and scRNA-seq

TSO concatemers of raw paired-end and single-end sequencing reads were counted with a custom bash script. A TSO concatemer was classified as any read containing more than one occurrence of the TSO sequence or its reverse complement. Raw sequencing reads were trimmed with fastp (v0.23.2) to remove sequencing adapters, TSO sequences, and G polymers. The trimmed reads were mapped with STAR (v2.7.11b) to the human reference genome GRCh38 (GENCODE 47, GRCh38 primary assembly) and subsequently uniquely mapped reads were quantified for each gene using featureCounts (v2.0.6) in paired-end mode without allowing overlap using the following parameters: `-p -countReadPairs -s 0 -t exon -g gene_id`. To normalize for gene length, the resulting raw counts per gene were transformed into TPM using a custom R script. Pairwise correlations between replicates and samples were calculated with  $\log_{10}(\text{TPM}+1)$ . Genes were considered as ‘detected’ with at least one count in one replicate and sample and considered as expressed with  $\text{TPM} \geq 1$ , respectively. ‘Gene body coverage’ was calculated using QoRTs (v1.3.6). ‘Percentages of mitochondrial reads’ were calculated by dividing the number of unique counts by the sum of total counts per sample. To schematize the mitochondrial chromosome, the package karyoploteR (1.24.0) was used and the corresponding mitochondrial genes symbols along with their genomic positions were added by using the R package biomaRt (v2.54.0). To visualize read coverage, samtools (v1.13) depth was used to obtain the read depth at each base for needed genomic ranges. Gene models were created with the R package ggtranscript (1.0.0) based on the GENCODE 47 annotation file. PCA was computed using Seurat’s (v5.2.1) RunPCA() function after filtering out ribosomal and mitochondrial reads. The counts were normalized with Seurat’s default function NormalizeData() and the 2000 most highly variable features were detected. After scaling the data, PCA was run on the highly variable genes with a number of 10 principal components. Furthermore, the RSeQC (v5.0.1) software package was used to examine the distribution of reads across different genomic features (CDS, 5’UTR, 3’UTR, introns, TSS\_up\_10kb, TES\_down\_10kb). In order to investigate the distribution of reads in more detail, the underlying RSeQC script read\_distribution.py was modified to extract each read and its corresponding read information (genomic start and end position, CIGAR length, read sequence) of each genomic element. Downstream analysis was conducted using R (v4.2.3) and Rstudio (v2022.07.2). The overlaps of reads mapping to genetic elements between uMRT and Smart-seq were calculated by generating genomic interval objects with a permissive window of 40 bp with the Genomic Ranges (v1.50.0) package. For sequencing saturation analysis and to account for the effect of different sequencing depths, reads were downsampled with seqtk (v1.4).

### **scSLAM-seq analysis**

The Quant-Seq, Smart-seq, and uMRT data were processed using the GRAND-SLAM pipeline (v3.0.7) (Jürges, Dölken and Erhard, 2018). Aligned reads were converted and merged using GEDI (Bam2CIT, and MergeCIT) (v1.0.6.c) (Erhard *et al.*, 2018). GRAND3 was then used (-auto -profile) to generate read counts and new-to-total ratios (NTRs) at the gene level as well as global mismatch statistics.

### **Data and code availability**

RNA-seq data generated in this study and data scripts that support the analysis in this manuscript are available on Zenodo: [10.5281/zenodo.17099015](https://zenodo.org/record/17099015)

### **Conflict of interest**

L.T.G., A.M.P., and A.-E.S. are cofounders of RNACoconnect, a company that is developing MRT as a reagent for sequencing and other applications. L.T.G. and A.M.P. are inventors on patents filed by Yale University for applications and adaptations of the MRT enzyme.

### **Acknowledgments**

We thank Brenton Graveley for advice and discussions on the experiments. This work was supported by National Institutes of Health (1R01HG011868). A.M.L., F.E. and A.-E.S. thank the Deutsche Forschungsgemeinschaft (DFG) Collaborative Research Center 1583 DECIDE (Project Z2, DFG project number: 492620490). A.E.S thanks the Agence Nationale pour la Recherche (ANR)-DFG support MAID (#530098272). A.G. thanks the Munich School of Data Science (MUDS). C.-L.C. thanks the HIRI Graduate Program “RNA and Infection”.

### **Author Contributions Statement**

Conceptualization, A.M.P. and A.-E.S.; Methodology, C.-L.C., and L.T.G.; Software/data analysis, A.G., T.R., and F.E.; supervision, A.M.P. and A.-E.S.; writing – original draft, A.-E.S.; writing – review & editing, All co-authors.

## References

- Boivin, V. *et al.* (2018) 'Simultaneous sequencing of coding and noncoding RNA reveals a human transcriptome dominated by a small number of highly expressed noncoding genes', *RNA*, 24(7), pp. 950–965.
- Carninci, P. *et al.* (1998) 'Thermostabilization and thermoactivation of thermolabile enzymes by trehalose and its application for the synthesis of full length cDNA', *Proceedings of the National Academy of Sciences of the United States of America*, 95(2), pp. 520–524.
- Chan, E.W. *et al.* (1980) 'Effects of polyethylene glycol on reverse transcriptase and other polymerase activities', *Biochimica et biophysica acta*, 606(2), pp. 353–361.
- Deng, Z., Chapagain, P. and Leng, F. (2023) 'Macromolecular crowding potently stimulates DNA supercoiling activity of Mycobacterium tuberculosis DNA gyrase', *The Journal of biological chemistry*, 299(12), p. 105439.
- Erhard, F. *et al.* (2018) 'Improved Ribo-seq enables identification of cryptic translation events', *Nature methods*, 15(5), pp. 363–366.
- Erhard, F. *et al.* (2019) 'scSLAM-seq reveals core features of transcription dynamics in single cells', *Nature*, 571(7765), pp. 419–423.
- Erhard, F. *et al.* (2022) 'Time-resolved single-cell RNA-seq using metabolic RNA labelling', *Nature reviews. Methods primers*, 2(1), pp. 1–18.
- González-Delgado, A. *et al.* (2021) 'Prokaryotic reverse transcriptases: from retroelements to specialized defense systems', *FEMS microbiology reviews*, 45(6). Available at: <https://doi.org/10.1093/femsre/fuab025>.
- Guo, L.-T. *et al.* (2022) 'Direct tracking of reverse-transcriptase speed and template sensitivity: implications for sequencing and analysis of long RNA molecules', *Nucleic acids research*, 50(12), pp. 6980–6989.
- Guo, L.-T. *et al.* (2024) 'Characterization and implementation of the MarathonRT template-switching reaction to expand the capabilities of RNA-seq', *RNA*, 30(11), pp. 1495–1512.
- Hagemann-Jensen, M. *et al.* (2020) 'Single-cell RNA counting at allele and isoform resolution using Smart-seq3', *Nature biotechnology*, 38(6), pp. 708–714.
- Hagemann-Jensen, M., Ziegenhain, C. and Sandberg, R. (2022) 'Scalable single-cell RNA sequencing from full transcripts with Smart-seq3xpress', *Nature biotechnology*, 40(10), pp. 1452–1457.
- Hahaut, V. *et al.* (2022) 'Fast and highly sensitive full-length single-cell RNA sequencing using FLASH-seq', *Nature biotechnology*, 40(10), pp. 1447–1451.
- Herzog, V.A. *et al.* (2017) 'Thiol-linked alkylation of RNA to assess expression dynamics', *Nature methods*, 14(12), pp. 1198–1204.
- Jedynak-Slyvka, M., Jabczynska, A. and Szczesny, R.J. (2021) 'Human mitochondrial RNA processing and modifications: Overview', *International journal of molecular sciences*, 22(15), p. 7999.
- Jürges, C., Dölken, L. and Erhard, F. (2018) 'Dissecting newly transcribed and old RNA using GRAND-SLAM', *Bioinformatics (Oxford, England)*, 34(13), pp. i218–i226.
- Kapteyn, J. *et al.* (2010) 'Incorporation of non-natural nucleotides into template-switching oligonucleotides reduces background and improves cDNA synthesis from very small RNA samples', *BMC genomics*, 11, p. 413.
- Kramer, R.M. *et al.* (2012) 'Toward a Molecular Understanding of Protein Solubility: Increased Negative Surface Charge Correlates with Increased Solubility' (2012) *Biophysical Journal*, 102(8), pp. 1907–1915.
- Kulpa, D., Topping, R. and Telesnitsky, A. (1997) 'Determination of the site of first strand transfer during Moloney murine leukemia virus reverse transcription and identification of strand transfer-associated reverse transcriptase errors', *The EMBO journal*, 16(4), pp. 856–865.
- Ludwig, L.S. *et al.* (2019) 'Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics', *Cell*, 176(6), pp. 1325–1339.e22.
- Martín-Alonso, S., Frutos-Beltrán, E. and Menéndez-Arias, L. (2021) 'Reverse Transcriptase: From Transcriptomics to Genome Editing', *Trends in biotechnology*, 39(2), pp. 194–210.
- Matz, M. *et al.* (1999) 'Amplification of cDNA ends based on template-switching effect and step-out PCR', *Nucleic acids research*, 27(6), pp. 1558–1560.

- Mereu, E. *et al.* (2020) 'Benchmarking single-cell RNA-sequencing protocols for cell atlas projects', *Nature biotechnology*, 38(6), pp. 747–755.
- Miller, T.E. *et al.* (2022) 'Mitochondrial variant enrichment from high-throughput single-cell RNA sequencing resolves clonal populations', *Nature biotechnology*, 40(7), pp. 1030–1034.
- Mohr, S. *et al.* (2013) 'Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing', *RNA (New York, N.Y.)*, 19(7), pp. 958–970.
- Monroy-Eklund, A. *et al.* (2023) 'Structural analysis of MALAT1 long noncoding RNA in cells and in evolution', *RNA (New York, N.Y.)*, 29(5), pp. 691–704.
- Moran, J.C. *et al.* (2024) 'The human mitochondrial mRNA structurome reveals mechanisms of gene expression', *Science (New York, N.Y.)*, 385(6706), p. eadm9238.
- Nottingham, R.M. *et al.* (2016) 'RNA-seq of human reference RNA samples using a thermostable group II intron reverse transcriptase', *RNA (New York, N.Y.)*, 22(4), pp. 597–613.
- Picelli, S. *et al.* (2013) 'Smart-seq2 for sensitive full-length transcriptome profiling in single cells', *Nature Methods*, 10(11), pp. 1096–1098.
- Quake, S.R. (2021) 'The cell as a bag of RNA', *Trends in genetics*, 37(12), pp. 1064–1068.
- Ramsköld, D. *et al.* (2012) 'Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells', *Nature biotechnology*, 30(8), pp. 777–782.
- Regev, A. *et al.* (2017) 'The Human Cell Atlas', *eLife*, 6. Available at: <https://doi.org/10.7554/eLife.27041>.
- Saliba, A.-E. *et al.* (2014) 'Single-cell RNA-seq: advances and future challenges', *Nucleic acids research*, 42(14), pp. 8845–8860.
- Santoro, M.M. *et al.* (1992) 'Increased thermal stability of proteins in the presence of naturally occurring osmolytes', *Biochemistry*, 31(23), pp. 5278–5283.
- Stark, R., Grzelak, M. and Hadfield, J. (2019) 'RNA sequencing: the teenage years', *Nature reviews. Genetics*, 20(11), pp. 631–656.
- Svensson, V., Vento-Tormo, R. and Teichmann, S.A. (2018) 'Exponential scaling of single-cell RNA-seq in the past decade', *Nature protocols*, 13(4), pp. 599–604.
- Tanay, A. and Regev, A. (2017) 'Scaling single-cell genomics from phenomenology to mechanism', *Nature*, 541(7637), pp. 331–338.
- Unlu, I. *et al.* (2024) 'Induro-RT mediated circRNA-sequencing (IMCR-seq) enables comprehensive profiling of full-length and long circular RNAs from low input total RNA', *Nucleic acids research*, 52(13), p. e55.
- Verwilt, J., Mestdagh, P. and Vandesompele, J. (2023) 'Artifacts and biases of the reverse transcription reaction in RNA sequencing', *RNA (New York, N.Y.)*, 29(7), pp. 889–897.
- Zhao, C., Liu, F. and Pyle, A.M. (2018) 'An ultraprocessive, accurate reverse transcriptase encoded by a metazoan group II intron', *RNA*, 24(2), p. 183.
- Zheng, G.X.Y. *et al.* (2017) 'Massively parallel digital transcriptional profiling of single cells', *Nature Communications*, 8(1), pp. 1–12.
- Zhu, Y.Y. *et al.* (2001) 'Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction', *BioTechniques*, 30(4), pp. 892–897.
- Zucha, D. *et al.* (2020) 'Performance Comparison of Reverse Transcriptases for Single-Cell Studies', *Clinical chemistry*, 66(1), pp. 217–228.
- Yao, Jun, Hengyi Xu, Elizabeth A. Ferrick-Kiddie, Ryan M. Nottingham, Douglas C. Wu, Manuel Ares Jr, and Alan M. Lambowitz. 2024. "Human Cells Contain Myriad Excised Linear Intron RNAs with Links to Gene Regulation and Potential Utility as Biomarkers." *PLoS Genetics* 20 (9): e1011416.