SMARTer single cell total RNA sequencing

Karen Verboom [1,2,†], **Celine Everaert**[1,2,†], **Nathalie Bolduc**[3], **Kenneth J. Livak**[4], **Nurten Yigit**[1,2], **Dries Rombaut**[1,2], **Jasper Anckaert**[1,2], **Simon Lee**[3], **Morten T. Venø**[5], **Jørgen Kjems**[5], **Frank Speleman**[1,2], **Pieter Mestdagh**[1,2] and **Jo Vandesompele** [1,2,*]

[1]Center for Medical Genetics, Ghent University, Ghent, Belgium, [2]Cancer Research Institute Ghent, Ghent, Belgium, [3]Takara Bio USA, Mountain View, CA 94043, USA, [4]Fluidigm Corporation, South San Francisco, CA 94080, USA and [5]Department of Molecular Biology and Genetics and Interdisciplinary Nanoscience Center, Aarhus University, Aarhus DK-8000, Denmark

ABSTRACT

Single cell RNA sequencing methods have been increasingly used to understand cellular heterogeneity. Nevertheless, most of these methods suffer from one or more limitations, such as focusing only on polyadenylated RNA, sequencing of only the 3′ end of the transcript, an exuberant fraction of reads mapping to ribosomal RNA, and the unstranded nature of the sequencing data. Here, we developed a novel single cell strand-specific total RNA library preparation method addressing all the aforementioned shortcomings. Our method was validated on a microfluidics system using three different cancer cell lines undergoing a chemical or genetic perturbation and on two other cancer cell lines sorted in microplates. We demonstrate that our total RNA-seq method detects an equal or higher number of genes compared to classic polyA[+] RNA-seq, including novel and non-polyadenylated genes. The obtained RNA expression patterns also recapitulate the expected biological signal. Inherent to total RNA-seq, our method is also able to detect circular RNAs. Taken together, SMARTer single cell total RNA sequencing is very well suited for any single cell sequencing experiment in which transcript level information is needed beyond polyadenylated genes.

INTRODUCTION

To understand the complexity of life, knowledge of cells as fundamental units is key. Recently, technological advances have emerged to enable single cell RNA sequencing (RNA-seq). In 2009, Tang *et al.* published the first single cell RNA-seq protocol in which cells were picked manually and transcripts reverse transcribed using a polydT primer (1). As the throughput was low, new methods using early multiplexing, such as STRT-seq and SCRB-seq, were introduced in which cells were pooled at an early step in the workflow, enabling processing of many cells in parallel (2–4). In contrast to these methods that have inherent 3′ end or 5′ end bias, Smart-seq2 generates read coverage across the whole transcript expanding the spectrum of applications as this method can be used for fusion detection, single nucleotide variants (SNV) analysis and splicing, beyond typical gene expression profiling applications (5,6). To reduce the polymerase chain reaction (PCR) bias generated in the aforementioned methods, CEL-seq and MARS-seq were introduced using linear in vitro transcription (IVT) instead of PCR to obtain enough cDNA for sequencing (7–9). Most recently, droplet and split-pool ligation based methods capturing thousands of single cells were developed, providing new insights in cellular heterogeneity and rare cell types (10–14). The main drawback of these methods is that analyses are typically confined to gene expression of only (3′ ends of) polyadenylated transcripts (Table 1).

More complex analyses with respect to alternative splicing, allele specific expression, mutation analysis, assembly of (novel) transcripts, circular RNA (circRNA) quantification and post-transcriptional regulation, require full-length and full-transcriptome methods. Moreover, sequencing a large number of cells is often compromising sequencing depth, resulting in low coverage per cell and detection of only the most abundant transcripts (15). In contrast to these droplet-based methods, microfluidic chip and flowcytometry based platforms typically capture fewer cells, but are able to sequence entire transcripts and detect a substantially higher number of genes per cell, providing a more complete view of the complexity and richness of single cells' transcriptomes (6,8). Of note, most single cell RNA-seq studies assess only 3′ end polyadenylated (polyA[+]) transcripts, ignoring non-polyadenylated (polyA[–]) transcripts (Table 1) (6,12,14). Since a substantial part of the human transcrip-

Table 1. Characteristics of the top ten cited single cell polyA[+] RNA-seq in Web of Science and four available single cell total RNA-seq methods (including our SMARTer method)

	Total RNA-seq	Full length	rRNA < 5%	Stranded	Reference
Drop-seq	−	−	+	−	(45)
Tang *et al.*	−	+	+	−	(1)
InDrop	−	−	+	−	(46)
MARS-seq	−	−	+	−	(9)
Smart-seq2	−	+	+	−	(6,40)
CEL-seq	−	−	+	+	(7)
STRT-seq	−	−	+	+	(3)
Quartz-seq	−	+	+	−	(47)
CEL-seq2	−	−	+	+	(8)
cytoSeq	−	−	+	−	(48)
SuPeR-seq	+	+	+	−	(19)
RamDA-seq	+	+	−	−	(20)
MATQ-seq	+	+	NA	−	(21)
SMARTer	+	+	+	+	

tome is non-polyadenylated, various RNA types, including circRNAs, enhancer RNAs, histone RNAs, and a sizable fraction of long non-coding RNAs (lncRNAs), are not quantified using these classic methods (16–18). In order to study polyA[–] transcripts at the single cell level, total RNA-seq workflows were developed (19–21). While in principle both polyA[+] and polyA[–] transcripts are converted into a sequencing-ready library using random primer mediated reverse transcription, these methods suffer from one or more of the following limitations: the strand-orientation information is lost and a high percentage of reads map to ribosomal RNA (rRNA) (Table 1). Therefore, new methods circumventing these limitations are warranted. A rRNA depletion step is essential as up to 95% of the total RNA content in a mammalian cell consists of rRNA. Moreover, to discriminate sense and antisense overlapping transcripts, stranded sequencing is crucial; at least 38% of the annotated transcripts in cancer cells have antisense expression (22). Here, we developed a novel easy to use and efficient single cell total RNA-seq workflow based on the SMARTer Stranded Total RNA-Seq Kit - Pico Input Mammalian combining for the first time strandeness and effective removal of ribosomal cDNA (Table 1). We ported the method to Fluidigm's C1 single cell microfluidics instrument, and demonstrated that the method works equally well on FACS sorted cells in microplates. In total, 458 cells from five different human cancer cell lines in four experiments were sequenced with a total sequencing depth of 1528 million reads. Using our novel method, we consistently observe <3% of ribosomal reads and we detect >5360 genes by at least four reads, including novel genes, polyA[–] genes and circRNAs.

MATERIALS AND METHODS

Cell lines

The neuroblastoma cell line NGP, used for the C1 experiments, is a kind gift of Prof. R. Versteeg (Amsterdam, the Netherlands). Cells were maintained in RPMI-1640 medium (Life Technologies, 52400-025) supplemented with 10% fetal bovine serum (PAN Biotech, P30-3306), 1% of L-glutamine (Life Technologies, 15140-148) and 1% penicillin/streptomycin (Life Technologies, 15160-047) (referred to as complete medium) at 37°C in a 5% CO_2 at-

mosphere. Short tandem repeat genotyping was used to validate cell line authenticity prior to performing the described experiments and mycoplasma testing was done on a monthly basis. The A375 (ATCC CRL-1619) and Jurkat (clone E6.-1; ATCC TIB-152) cells, used for the FACS experiments, were grown in Dulbecco's modified Eagle's medium (DMEM; Millipore-Sigma, D5796) supplemented with 10% Tet system approved fetal bovine serum (FBS) (Takara, 631106) and RPMI-1640 medium (RPMI; Millipore-Sigma, R0883) supplemented with 10% Tet system approved FBS, respectively. Cell lines were sub-cultured every two days or when they reached >80% confluence (A375) or >1 × 10^6 cells/ml (Jurkat).

Cell cycle synchronization and nutlin-3 treatment of NGP cells

NGP cells were synchronized using serum starvation prior to nutlin-3 treatment. First, cells were seeded at low density for 48 h in complete medium. Then, cells were refreshed with serum-free medium for 24 h. Finally, the cells were treated with either 8 μM of nutlin-3 (Cayman Chemicals, 10004372, dissolved in ethanol) or vehicle. Cells were trypsinized (Gibco, 25300054) 24 h post treatment and harvested for single cell analysis, bulk RNA isolation and cell cycle analysis.

Cell cycle analysis

Four million cells were washed with PBS (Gibco, 14190094) and the pellet was resuspended in 300 μl PBS. Next, 700 μl of 70% ice-cold ethanol was added dropwise while vortexing to fix the cells. Cells were stored at −20°C for at least 1 h. After incubation, cells were washed with PBS and the pellet was resuspended in 1 ml PBS containing RNAse A (Qiagen, 19101) at a final concentration of 0.2 mg/ml. After 1 h incubation at 37°C, propidium iodide (BD biosciences, 556463) was added to a final concentration of 40 μg/ml. Samples were loaded on a S3 cell sorter (Bio-Rad) and analyzed using the FlowJo v.10 software.

RNA isolation and cDNA synthesis

Total RNA was isolated using the miRNeasy mini kit (Qiagen, 217084) with DNA digestion on-column accord-

ing to the manufacturer's instructions. RNA concentration was measured using spectrophotometry (Nanodrop 1000, Thermo Fisher Scientific). cDNA was synthesized using the iScript Advanced cDNA synthesis kit (Bio-Rad, 1708897) using 500 ng RNA as input in a 20 µl reaction. cDNA was diluted to 2.5 ng/µl with nuclease-free water prior to RT-qPCR measurements.

Reverse transcription quantitative PCR

PCR mixes containing 2.5 µl 2× SsoAdvansed SYBR qPCR supermix (Bio-Rad, 04887352001), 0.25 µl each forward and reverse primer (5 µM, IDT), and 2 µl diluted cDNA (5 ng total RNA equivalents) were analyzed on the LightCycler480 instrument (Roche) using two replicates. Expression levels were normalized using expression data of four stable reference genes (*SDHA, YWHAZ, TBP, HPRT1*). RT-qPCR data was analyzed using the qbase+ software v3.0 (Biogazelle). Primer sequences are available in Supplementary Table S1.

FACS sorting of A375 and Jurkat cells in microplates

Before sorting, cells were washed twice in 1× PBS buffer (DPBS without calcium chloride and magnesium chloride; Sigma Aldrich, D8537) and labelled with 7-AAD (BD Pharmingen, 51-68981E) for live/dead differentiation and FITC-conjugated antibody [anti-CD47 (BD Pharmingen, 556045) for A375 and anti-CD81 (BD Pharmingen, 551108) for Jurkat]. After washing off the unbound antibodies in 1× PBS, cells were resuspended in BD FACS Pre-Sort Buffer (BD, 563503). Single cell sorting in 8-tube PCR strips was done using a BD FACSJazz Cell Sorter. A375 cells were sorted in 7 µl 1× PBS buffer and Jurkat cells in 8 µl lysis solution [100 µl 10× Lysis buffer (Takara, 635013), 5 µl RNase Inhibitor (Takara, 635013) and 700 µl water]. Following sorting, tubes were sealed and subjected to a quick spin and immediately frozen on dry ice and finally stored at −80°C until use. All sorting experiments included negative controls (no cell in a well).

Single cell total RNA library preparation of nutlin-3 treated NGP cells

Cells were washed with PBS and pellets of vehicle treated cells were resuspended and incubated in 1 ml pre-warmed (37°C) cell tracker (CellTracker Green BODIPY Dye, Thermo fisher Scientific, C2102) for 20 min at room temperature. After incubation, cells were washed in PBS and resuspended in 1 ml wash buffer (Fluidigm, 100-6201). An equal number of stained (vehicle treated) and non-stained (nutlin-3 treated) cells were mixed and diluted to 300 000 cells/ml. Suspension buffer (Fluidigm) was added to the cells in a 3:2 ratio and 6 µl of this mix of was loaded on a primed C1 Single-Cell Open App IFC (Fluidigm, 100-8134) designed for medium-sized cells (10–17 µm). Cells were captured using the 'SMARTer single cell total RNA-seq' script deposited in Script Hub (Fluidigm). Upon capture, cells were visualized using the Axio Observer Z1 (Zeiss) and a median multiplet rate of 34.54% was observed over all experiments. These cells were excluded from further analyses. Sequencing libraries were generated using the C1 running the

'SMARTer single cell total RNA-seq' script deposited on Script Hub. In short, the SMARTer Stranded Total RNA-Seq Kit v2 - Pico Input Mammalian (Pico v2, total RNA, Takara, 634413) was used to synthesize cDNA with following modifications. Cells were fragmented and lysed by loading 7 µl of 10× reaction mix [2.3 µl SMART Pico Oligo Mix v2, 6 µl 5× first-strand buffer, 1 µl 20× C1 loading reagent (Fluidigm), 3 µl lysis mix (19 µl 10× lysis buffer, 1 µl RNAse inhibitor (40 U/µl)), 1 µl 1/1000 diluted ERCC spikes (Ambion, 4456740), 6.7 µl water] and incubating the cells at 85°C for 6 minutes (to lyse cells and fragment RNA) followed by 2 min at 10°C. Next, 8 µl first strand master mix [1 µl C1 loading reagent, 4 µl 5× first-strand buffer, 0.9 µl RNAse inhibitor (40 U/µl), 3.5 µl SMARTScribe reverse transcriptase (100 U/µl), 7.9 µl SMART TSO Mix v2 (from Takara kit, 634413), 2.7 µl water] was loaded and incubated at 42°C for 90 min followed by 70°C for 10 min. Finally, a PCR master mix for each well was made [1 µl 20× loading reagent, 2 µl 2.4 µM forward primer (Takara, 634413), 2 µl 2.4 µM reverse primer, 13.1 µl 1.5× PCR mix (1050 µl 2× SeqAmp CB buffer, 42 µl SeqAmp DNA polymerase, 308 µl water)] and 5 µl of each of these mixes was loaded in the harvest wells of the IFC. The samples were incubated for 1 min at 94°C followed by 11 PCR cycles (30 s at 98°C, 15 s at 55°C, 30 s at 68°C) and 2 min at 68°C. Following this initial cDNA amplification, 12 wells were pooled per tube using 8 µl of cDNA per cell. Next steps of the library prep were performed according to manufacturer's instructions with minor modifications. 13 PCR cycles were used for PCR2 and a 1:1 ratio was used for beads cleanup after PCR2. Next, the samples were resuspended in 22 µl 5 mM tris buffer (from kit) and 20 µl was used to perform a second beads cleanup using a 0.9:1 ratio. Finally, the samples were resuspended in 12 µl tris buffer and the quality was determined on the Fragment Analyzer (Advanced Analytical). Of note, the protocol can also be executed using the single cell specific version of the kit, released by Takara (SMART-Seq Stranded Kit, 634442) after we had completed our C1 experiments.

Single cell polyA[+] RNA library preparation of nutlin-3 treated NGP cells

Vehicle treated cells were stained with cell tracker as described above. An equal number of stained (vehicle treated) and non-stained (nutlin-3 treated) cells were mixed and diluted to 300 000 cells/ml. Suspension buffer was added to the cells in a 3:2 ratio and 6 µl of this mix of was loaded on a primed C1 Single-Cell Auto Prep Array for mRNA Seq (Fluidigm, 100-6041) designed for medium-sized cells (10–17 µm). Single cell polyA[+] RNA sequencing on the C1 was performed using the SMART-Seq v4 Ultra Low Input RNA Kit for the Fluidigm C1 System (SMART-Seq v4, polyA[+] RNA, Takara, 635026) according to manufacturer's instructions. One microliter of the ERCC spike-in mix was diluted in 999 µl loading buffer to get a 1/1000 dilution of the ERCC spikes. One microliter of this dilution was added to the 20 µl lysis mix. The quality of the cDNA was checked for 11 random single cells on the Fragment Analyzer. The concentration of the cells was measured using the quantifluor dsDNA kit (Promega, E2670) and glo-

max (Promega) according to manufacturer's instructions. The samples were 1/5 diluted in C1 harvest reagent (Fluidigm). Next, library prep was performed using the Nextera XT library prep kit (Illumina, FC-131-1096) according to manufacturer's instructions, followed by quality control on the Fragment Analyzer.

Single cell total RNA library preparation of FACS sorted A375 and Jurkat cells

Cells were processed using the SMARTer Stranded Total RNA-Seq Kit v2 – Pico Input Mammalian (Takara, 634413) or the SMART-Seq Stranded Kit (Takara, 634444) reagents according to the manufacturer's instructions with some modifications that were also implemented in the C1 protocol. For the SMART-Seq Stranded Kit, the Ultra Low Input workflow described in the user manual was followed by pooling of eight samples according to Appendix A of the user manual. For the SMARTer Stranded Total RNA-Seq Kit v2 – Pico Input Mammalian, the cells were also processed as described for the SMART-Seq Stranded Kit, but using the reagents specific to the SMARTer Stranded Total RNA-Seq Kit v2 – Pico Input Mammalian, which were also used for the C1 protocol. For both kits, cells sorted in a lysis solution instead of $1\times$ PBS were processed without addition of lysis buffer. Identital to the C1 protocol, the initial RNA shearing step was performed at $85°C$ for 6 min and 10 and 13 PCR cycles were carried out for PCR1 and PCR2, respectively.

Library sequencing

All libraries were quantified using the KAPA library quantification kit (Roche) and libraries were diluted to 4 nM. For NGP, the polyA[+] RNA library and total RNA library were pooled in a 1/4 ratio and 1.5 pM of the pooled library was single-end sequenced on a NextSeq 500 (Illumina) with a read length of 75 bp and a total sequencing read depth of 274 million reads, combining single cell polyA[+] and total RNA libraries to prevent inter-run bias. A median sequencing read depth of 0.81 and 3.67 million reads per cell was reached for the single cell polyA[+] and total RNA libraries, respectively. In addition, 1.3 pM of the total RNA library was also sequenced in 2×75 paired-end sequencing run mode on the NextSeq 500, yielding 327 million reads and a median sequencing read depth of and 4.04 million per cell. The fastq data is deposited in GEO (GSE119984). A375 and Jurkat total RNA libraries were pooled and 1.2 pM of the pooled library was sequenced in 2×75 paired-end run mode on the NextSeq 500, yielding 41 million reads. FASTQ data is deposited in GEO (GSE130578).

Sequencing data quality control

While single-end sequencing libraries do not require pre-trimming, the paired-end libraries were trimmed using cutadapt (v.1.16) (23) to remove three nucleotides of the 5′ end of read 2. To assess the quality of the data, the reads were mapped using STAR (v.2.5.3) (24) on the hg38 genome including the full ribosomal DNA (45S, 5.8S and 5S) and mitochondrial DNA sequences. The parameters of STAR

were set to retain only primary mapping reads, meaning that for multi-mapping reads only the best scoring location is retained. Using SAMtools (v1.6) (25), reads mapping to the different nuclear chromosomes, mitochondrial DNA and rRNA were extracted and annotated as exonic, intronic or intergenic. In contrast to the unstranded nature of polyA[+] Smart-seq v4 data, the total RNA SMARTerseq data is stranded and processed accordingly (unless explicitly mentioned). Gene body coverage was calculated using the full Ensembl (v91) (26) transcriptome. The coverage per percentile was calculated, followed by a loess regression fit.

### Quantification of Ensembl and LNCipedia genes

Genes were quantified by Kallisto (v.0.43.1) (27) using both Ensembl (v.91) (26) extended with the ERCC spike sequences and LNCipedia (v.5.0) (28). The strandedness of the total RNA-seq reads was considered by running the –rf-stranded mode and omitted for unstranded analysis of the data. Subsampling 1 million reads (polyA[+] RNA libraries) or 1, 4 or 8 million reads (total RNA libraries) was performed by seqTK (v.1.2) followed by Kallisto quantification. Further processing was done with R (v.3.5.1) making use of tidyverse (v.1.2.1). To measure the biological signal we first performed differential expression analysis between the treatment groups using DESeq2 (v.1.20.0) (29) in combination with Zinger (v.0.1.0) (30). To identify enriched gene sets a fsgea (v.1.6.0) analysis was performed, calculating enrichment for the hallmark gene sets retrieved from MSigDB (v.6.2).

Circular RNA detection

CircRNAs were detected using the deeper sequenced paired-end sequencing data. Trim_galore (v.0.4.1) was used to trim adaptor sequences, perform quality filtering and remove three nucleotides from the 5′ end of read 2. Subsequently, reads from all samples were combined, adding originating sample names to read names for later splitting of data. The combined data was used for circRNA detection using find_circ (v.1) (31) using the reads2sample (find_circ.py -r) option to allow circRNA detection on the combined dataset while dividing out the contribution from each sample in the output. Only circRNAs with unique mapping on both anchors were accepted. Human genome hg19 was used for circRNA analysis. CircRNAs were annotated with host gene names from RefSeq (release 75) and circBase IDs from circbase.org. The Database for Annotation, Visualisation and Integrated Discovery (DAVID, v.6.8) (32,33) was used for Gene Ontology (GO) analysis for the circRNA host genes using biological processes (BP) and molecular function (MF). P-value <0.05 was used for statistical significance.

Single cell transcriptome assembly

A transcriptome per cell was created by combining STAR (v.2.5.3) and Stringtie (v.1.3.0) (34), using the deeper sequenced paired-end sequencing data. The parameters of Stringtie were set to require a coverage of 1. These single cell
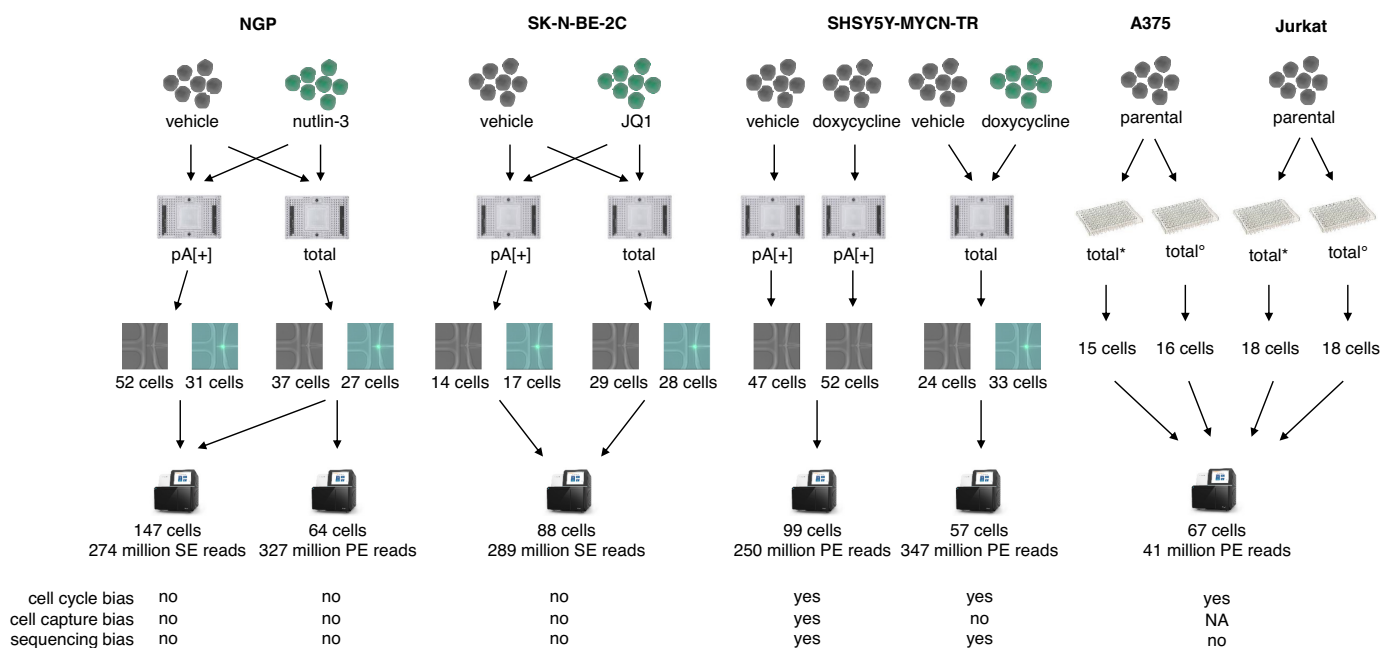
5 cell lines – 458 cells – 1528 million reads

Figure 1. Overview of experimental set-up. Single cell total RNA libraries of the FACS sorted cells were generated using two different reagent kits (#634413, denoted with * and #634444, denoted with °).

transcriptomes were merged with the Ensembl (v.91) transcriptome as a reference. From the merged multi-cell transcriptome, only multi-exonic genes with a minimum length of 200 nt were retained. To define the set of novel genes, genes annotated in Ensembl (26) or LNCipedia (v.5.0) (28) were filtered out. All genes in this novel multi-cell transcriptome were quantified using Kallisto on single-end subsampled data (1, 4 or 8 million reads per cell). Genes with an estimated count higher than 1 were retained.

RESULTS

Principle of SMARTer single cell total RNA sequencing

We developed a single cell total RNA-seq protocol for unbiased, full transcript and strand-specific analysis of both polyadenylated and non-polyadenylated transcripts from mammalian cells. The method uses reagents from the SMARTer Stranded Total RNA-Seq Kit v2 – Pico Input Mammalian (Pico v2, total RNA), a kit that is meant for low input bulk total RNA-seq, whereby we optimized reaction volumes, number of PCR cycles, and duration and temperature of the RNA fragmentation. The library preparation method employs random primers and a template switching mechanism to capture full transcript fragments of both polyadenylated (polyA[+]) and non-polyadenylated (polyA[−]) transcripts. Unwanted ribosomal cDNA is removed using probes, complementary to mammalian rRNA. After successfully porting the bulk library prep protocol to Fluidigm's C1 single cell instrument, we assessed the performance of the single cell total RNA-seq protocol through three distinct experiments in which nutlin-3, JQ1 or doxy-

cycline was used to treat NGP, SK-N-BE-2C and SHSY5Y-MYCN-TR neuroblastoma cell lines, respectively (with vehicle treated cells as control) (Figure 1). In addition, we performed matched single cell polyA[+] RNA-seq as a reference using cells from the same pool. While all experiments were successful, we focus our analyses and performance assessment on the NGP data. In this experiment, the treated and control cells were processed in the same microfluidic chip (preventing possible chip bias), the highest number of cells were captured, and the highest sequencing depth was reached.

SMARTer single cell total RNA sequencing yields high-quality data

In single cell sequencing experiments, it is important to prevent or limit potential biases that mask true biological differences. In particular, the cell cycle state is a known confounder (35). Therefore, we synchronized the cells through serum starvation for 24 hours. Upon synchronization, 80.3% of the NGP cells showed an arrest at the G0/G1 stage compared to only 53.3% for non-synchronized NGP cells (Supplementary Figure S1A and B). Subsequently, the synchronized NGP cells were treated for 24 h with vehicle or nutlin-3, the latter known to release TP53 from its negative regulator MDM2. As expected, nutlin-3 treatment resulted in cell cycle arrest (Supplementary Figure S1C and D). To prevent possible C1 batch effects (36), vehicle treated NGP cells were stained and loaded together with the non-stained nutlin-3 treated cells on the same C1 chip. Based on the fluorescent label and the transparency of the C1 system, vehicle and nutlin-3 treated cells were dis-
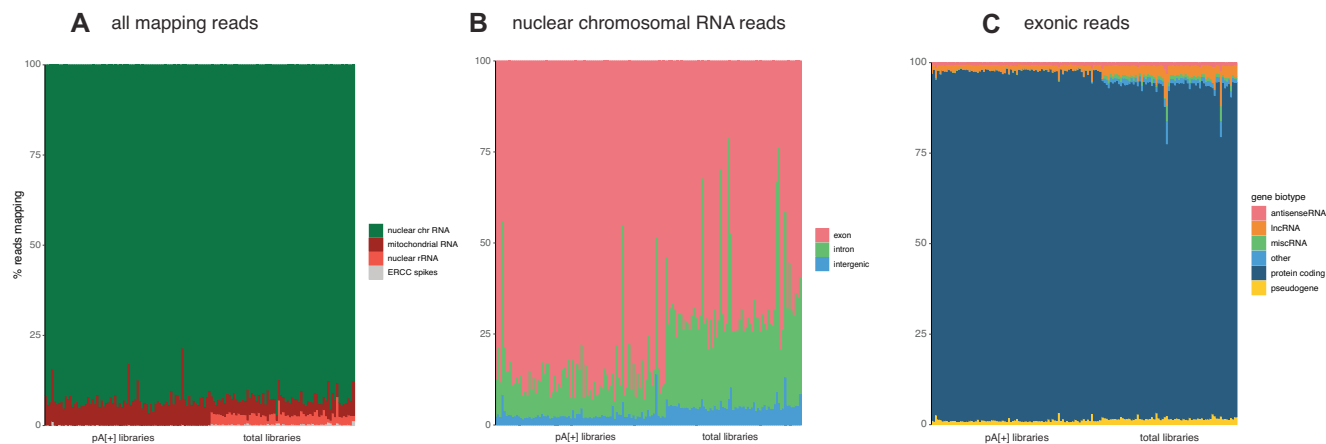
Figure 2. Read distribution differs between polyA[+] and total RNA libraries. (**A**) Percentage of reads derived from nuclear RNA, mitochondrial RNA and ribosomal RNA per cell quantified with STAR. (**B**) Percentage of the reads originating from nuclear chromosomes derived from exonic, intronic and intergenic regions per cell quantified with STAR. (**C**) Percentage of exonic reads attributed to the different biotypes per cell quantified with Kallisto.

criminated by fluorescence microscopy. By loading two C1 chips, one for polyA[+] RNA and one for total RNA library preparation, we captured 31 and 27 nutlin-3 treated versus 52 and 37 vehicle treated single cells, respectively. High-quality cDNA libraries of polyA[+] and total RNA were generated using the SMART-Seq v4 Ultra Low Input RNA Kit for the Fluidigm C1 System (SMART-Seq v4, polyA[+]) and our novel SMARTer single cell total RNA-seq protocol, respectively (Supplementary Figure S1E and F). ERCC spike-in molecules were added for external quality control in the lysis mix (Supplementary Figure S2). For the recovered spikes (with a concentration in the original mix of at least 10 amol/μl), linear models were calculated (Supplementary Figure S3), retrieving similar R^2 values for the polyA[+] RNA and total RNA library preparation protocol (Supplementary Figure S4). The transcripts detected in the polyA[+] libraries were somewhat shorter compared to the total RNA libraries (Supplementary Figure S5). In addition, the total RNA-seq libraries show a more uniform transcript coverage (Supplementary Figure S6).

As expected, a higher fraction of reads mapped to nuclear rRNA in the total RNA-seq libraries compared to the polyA[+] RNA libraries (average of 2.739% [2.488, 2.990; 95% confidence interval (CI)] versus 0.031% [0.026, 0.035; 95% CI], respectively). Nevertheless, the fraction of nuclear rRNA is very low in the total RNA libraries considering the use of random priming data (Figure 2A), and substantially lower compared to the RAMDA-seq method (9.667% rRNA [9.615, 9.719; 95% CI], Supplementary Figure S7). Furthermore, the single cell total RNA libraries contain more intronic (27.99% [25.06, 30.91; 95% CI] versus 11.87% [10.14, 13.60; 95% CI]) and intergenic (5.38% [5.00, 5.76; 95% CI] versus 2.90% [2.54, 3.26; 95% CI]) reads originating from nuclear chromosomes compared to polyA[+] RNA libraries (Figure 2B). Non-polyadenylated histone genes are highly abundant in the total RNA libraries, while low or absent in the polyA[+] libraries, confirming the validity of our single cell total RNA-seq workflow (Supplementary Figure S8). Equal results were obtained for the SK-N-BE-2C, and SHSY5Y-MYCN-TR cell lines (Supplementary Figure S9).

SMARTer single cell total RNA sequencing reveals a unique set of genes

More reads map to long intergenic RNAs (lincRNAs) using the single cell total RNA-seq protocol (2.64% [2.523, 2.756; 95% CI]) compared to polyA[+] RNA sequencing (1.67% [1.489, 1.849; 95% CI]). In addition, the single cell total RNA-seq protocol detects an equal or higher number of genes (subsampled to 1 million reads/cell and detected by >10 reads) covering the different biotypes, including lincRNAs (144 [139, 148; 95% CI]), protein coding (5124 [4874, 5372; 95% CI]) genes, and pseudogenes (132 [127, 137; 95% CI]) (Figures 2C and 3). Of note, antisense genes are the only biotype for which the total RNA protocol detects fewer genes (62 [59–64; 95% CI]), likely because of the unstranded nature of the polyA[+] RNA libraries, which results in erroneous quantification of sense/antisense overlapping genes (Supplementary Figure S10). Considering both polyA[+] RNA-seq and total RNA-seq data, 3978 different antisense-sense relationships with an overlap of >200 nucleotides were detected with expression of the sense or antisense gene in at least one cell. These loci are prone to erroneous quantification. Quantification of the stranded SMARTer data in an unstranded way shows that 42.1% (median of 180 of the 428 detected antisense genes per cell) of the detected antisense genes (in six random cells) are receiving counts, while they have zero counts when properly treated as stranded data; further, 10.1% of the antisense genes detected in both analyses display fold change differences larger than 2 (Supplementry Figure S11). Most of these genes with fold change differences (87.0%) are more abundant in the unstranded analysis compared to the stranded analysis, explained by the fact that these antisense genes are consuming counts from the sense gene.

LincRNAs, antisense genes and pseudogenes are clearly expressed in fewer cells compared to protein coding genes. We hypothesize that low abundant genes might be missed because of sampling bias during the sequencing workflow or that lincRNAs, often low abundant in nature, are expressed under specific conditions or stimuli (Figure 4) (37). As expected, increasing the number of reads (up to 4 or 8
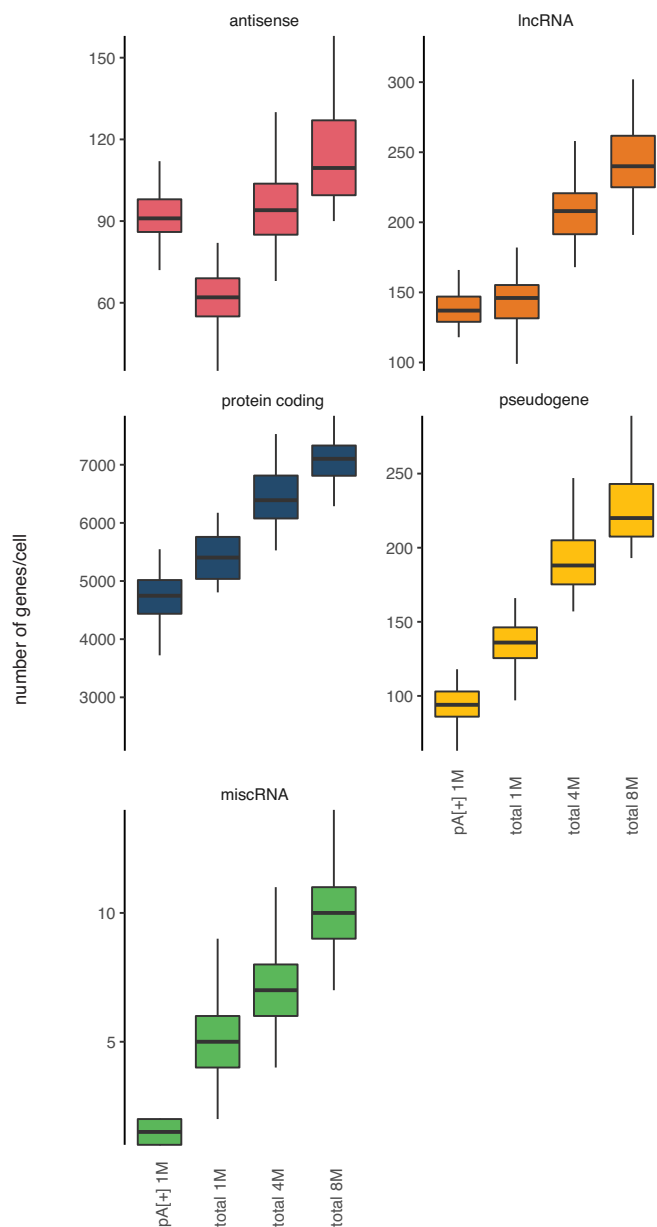
Figure 3. Total RNA libraries comprise more genes per RNA biotype. All genes in Ensembl v.91 were quantified on subsampled data (1, 4 or 8 million reads per cell). Only genes with at least 10 reads were included.
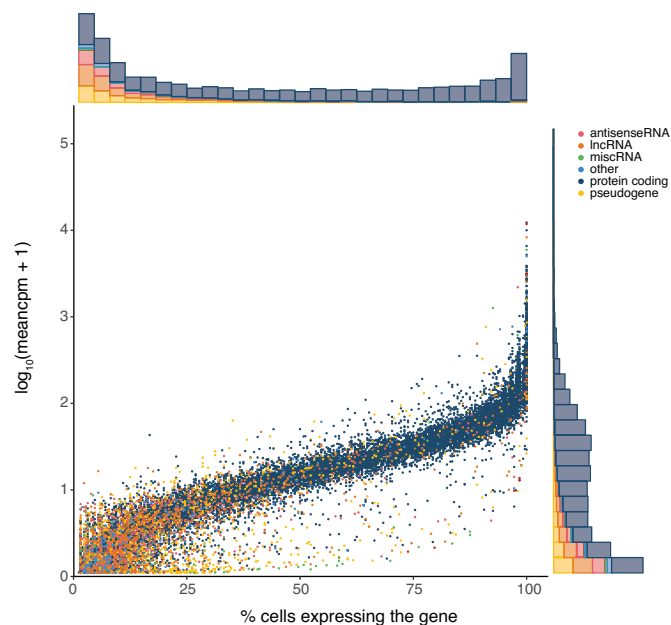

Figure 4. Gene biotype and abundance are correlated to fraction of expressed cells. In general, the fraction of cells in which a gene is expressed is related to the mean expression level of that gene; exceptionally, some low abundant genes are present in a large fraction of cells. RNA biotypes that are known to be more cell-type specifically expressed, such as lincRNAs, are expressed in fewer cells.

cific lncRNAs display a high expression, thus possibly representing functionally important RNAs (Figure 5D). LincRNA *RMRP* is one of the most abundant lincRNAs that is solely detected by our novel single cell total RNA-seq workflow. This gene is known to be 3′ non-adenylated and is the first known RNA encoded by a single-copy nuclear gene imported into mitochondria (38,39). As only a subset of the lincRNAs and antisense genes are currently annotated in Ensembl, we also quantified our libraries with the LNCipedia transcriptome (the most comprehensive human resource of both antisense and lincRNA genes, further referred to as lncRNAs). While the number of detected lncRNAs is slightly lower in the total RNA-seq libraries if an equal number of reads (1 million) is used, each library type contains a certain proportion of unique lncRNAs (Supplementary Figure S12). LNCipedia is likely biased towards medium-to-high abundant polyadenylated lncRNAs.

SMARTer single cell total RNA sequencing detects circular RNAs and novel genes

In addition to linear RNA biotypes, we tested whether the single cell total RNA-seq protocol is able to quantify circRNAs as this class of non-coding RNAs lacks a polyA-tail and in principle can only be detected using unbiased total RNA-seq. With a requirement of at least two unique back-spliced junction reads, 537 circRNAs were identified derived from 460 host genes (Supplementary Table S2). The majority of the circRNAs were found in fewer than 3 out of 64 cells, with only 14 circRNAs detected in at least four cells. Gene Ontology analysis for molecular functions and biological processes was performed on the circRNA host
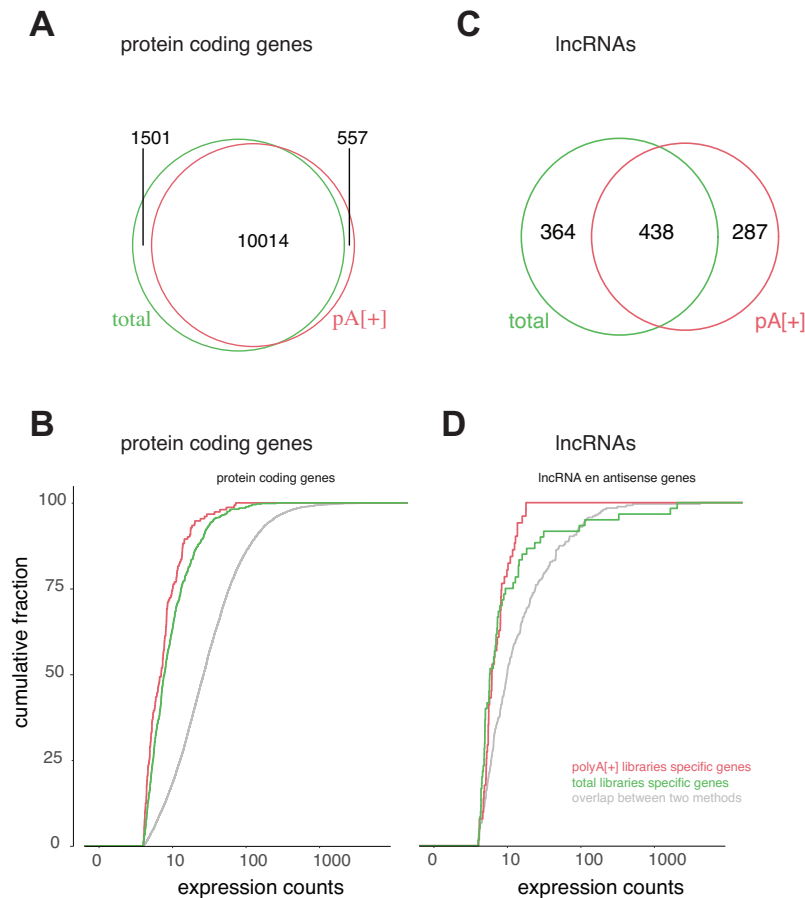
million) in the total RNA library protocol results in the detection of a higher number of genes. We observed no saturation when generating 8 million reads per cell, suggesting that deeper sequencing could yield even more detected genes (Figure 3). The overlap between protein coding genes detected in the polyA[+] and total RNA libraries (subsampled for 1 million reads/cell and mean expression of at least 1 read over all cells) (Figure 5A) is high. Genes detected in only one of the library types are generally lower abundant compared to genes detected with both methods (Figure 5B). In contrast to protein coding genes, the overlap for lincRNAs between the methods is much smaller (Figure 5C). Importantly, a significant fraction of the total RNA-seq spe-

Figure 5. While most protein coding genes are commonly detected, lincRNAs appear more method specific. (**A**) Overlap between protein coding genes detected in polyA[+] (1 million reads) and total RNA (1 million reads) libraries. (**B**) Expression counts for protein coding genes detected in only polyA[+] libraries (red), only total RNA libraries (green) or both (gray). (**C**) Overlap between lncRNAs detected in polyA[+] (1 million reads) and total RNA (1 million reads) libraries. (**D**) Expression counts for lncRNAs detected in only polyA[+] libraries (red), only total RNA libraries (green) or both (gray).

genes from both treated and untreated cells. A significant enrichment of TP53 binding, TP53 pathway, cell cycle, and chromosome organization suggests that the identified circRNAs may play a role in these biological functions.

In the single cell total RNA libraries, the fraction of intergenic reads (relative to existing Ensembl and LNCipedia annotation) is high, suggesting that these reads originate from novel unannotated transcripts. To validate this hypothesis, we generated genome and transcriptome guided transcriptome assembly of the paired-end single cell total RNA-seq data resulting in 5360 novel, multi-exonic genes. The novel transcripts have a median length of 317 nucleotides (Figure 6A) and consist on average of more than three exons (Figure 6B). Quantification of this novel transcriptome using the single-end data subsampled at 1 million reads per cell resulted in a median number of 59 novel genes per cell [55–63; 95% CI] (Figure 6C). Of note, most novel genes are expressed in only one cell (Figure 6D).

SMARTer single cell total RNA profiles reflect the biological signal

To assess whether the single cell total RNA-seq protocol is also able to reveal known biological signal, we performed

differential expression analysis using DESeq2 combined with the Zinger method coping with zero inflated data. Based on the ranking obtained by the DESeq2 test statistic, gene set enrichment analysis using the hallmark gene sets was performed. Firstly, the same gene sets are significantly enriched in both library preparation protocols (Figure 7A); secondly, TP53 target genes are—as expected—the most significantly enriched gene set (Figure 7B), confirming that the biological signal is recapitulated through single cell total RNA-seq analyses.

SMARTer single cell total RNA sequencing of FACS sorted cells in microplates

To demonstrate that our novel single cell total RNA seq method also efficiently works on FACS sorted cells in microplates, we processed A375 and Jurkat sorted cells. In parallel, Takara's single cell purposed SMART-Seq Stranded Kit was also tested on these cells (Figure 1). Equally low amounts of ribosomal cDNA were sequenced using both reagent kits, i.e. 1.46% [0.77, 2,15; 95% CI] and 0.66% [0.48, 0.85; 95% CI] for the A375 cells and 1.17% [1.05, 1.29; 95% CI] and 0.94% [0.80, 1.09; 95% CI] for the Jurkat cells (Figure 8, Supplementary Figure S13). Similar to the total RNA
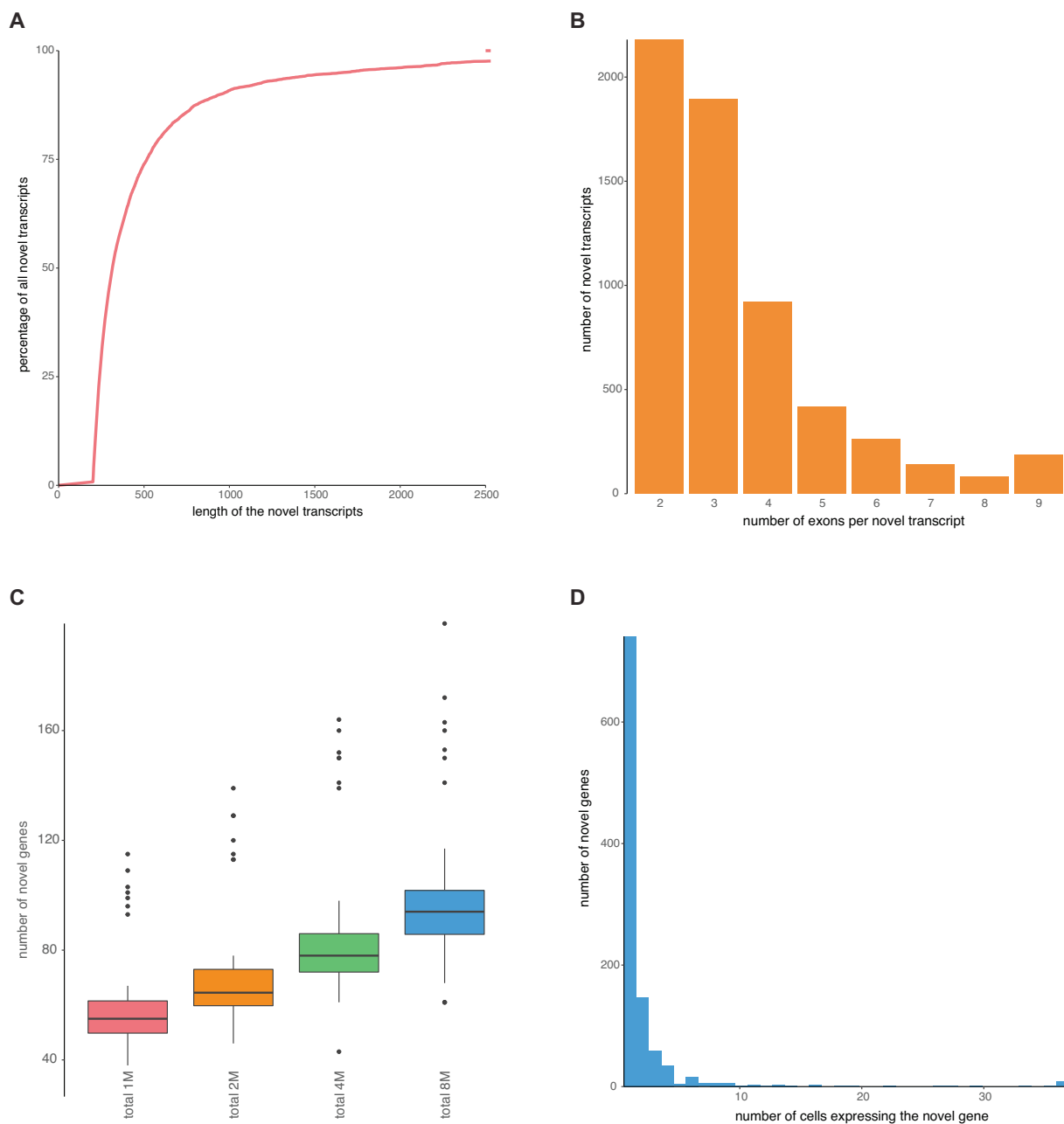
Figure 6. Total RNA libraries enable assembly of single cell transcriptomes. (**A**) Transcripts were filtered at a length of 200 nt. The remaining transcripts have a mean length of 537 nt. (**B**) Transcripts were required to have at least two exons. The remaining transcripts are on average 3.4 exons long. (**C**) All novel genes were quantified on subsampled data (1, 4 or 8 million single-end reads per cell). Genes with at least one count were retained. (**D**) While some novel genes are expressed in all cells, most novel genes are detected in only one cell.

seq libraries generated on the C1 system, we analysed the number of reads assigned to intron, exon and intergenic regions and the read fraction for all RNA biotypes. The microplate sorted single cell data was very comparable to the C1 data (Figure 8, Supplementary Figure S13).

DISCUSSION

In this study, we developed a single cell total RNA-seq method to sequence full transcripts from single cells in an essentially unbiased manner. To demonstrate the perfor-

mance of the method, we applied single cell total RNA-seq in four experiments on five different cancer cell lines, of which three undergoing a specific perturbation. In parallel, we also performed single cell polyA[+] RNA-seq on three cell lines using the well-established Smart-seq v4 method (6,40). As in any genomics study, the experimental set-up may suffer from confounding factors, such as variations in cell cycle states of the cells and batch effects of single cell capture and sequencing, masking real biological differences. In two of the four experiments, we carefully controlled all these experimental biases. The cell cycle bias was minimized
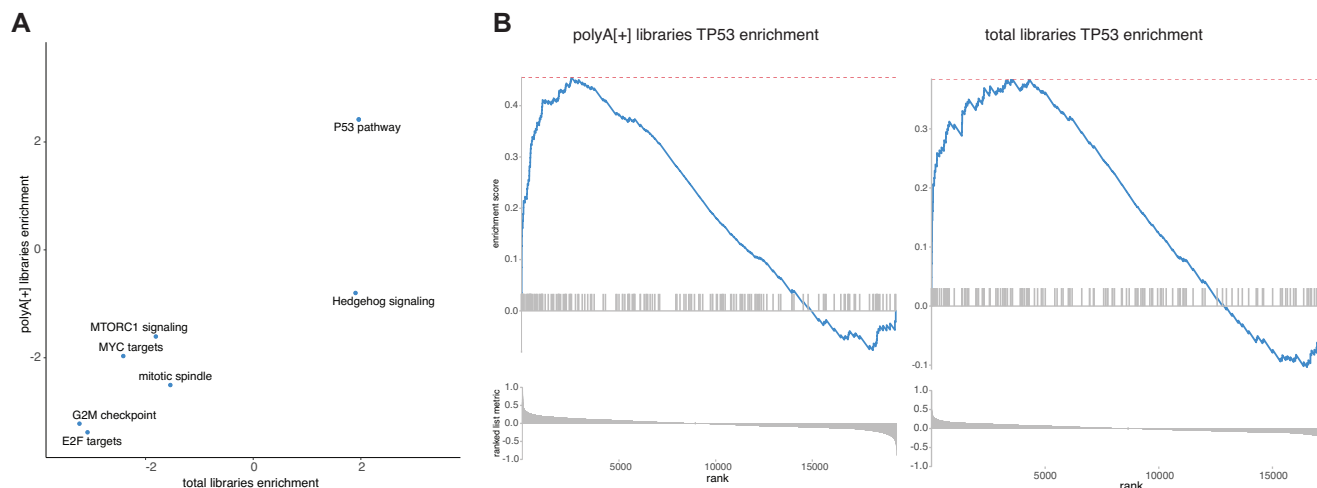
Figure 7. Pathway analysis for polyA[+] RNA and total RNA libraries is similar. (**A**) Gene set enrichment analysis for all hallmark pathways resulted in the same significant ($P_{adj} < 0.05$) pathway predictions. (**B**) The TP53 pathway is, as expected, enriched in both library prep methods.

by cell cycle synchronization using serum starvation. We also avoided potential cell selection bias by capturing differentially labeled treated and untreated cells on the same chip (35,36). Finally, sequencing bias was minimized by sequencing both polyA[+] and total RNA libraries on the same Illumina flow cells.

The single cell total RNA-seq method has some distinctive advantages compared to other methods. First, in any total RNA-seq library, depletion of rRNA is essential as this makes up the bulk of the total RNA mass. Depletion of rRNA from single cells prior to cDNA synthesis is technically very difficult. Here, we used ribosomal cDNA specific removal probes, resulting in <3% of ribosomal reads per single cell library. This highly efficient rRNA depletion step is a major improvement compared to RAMDA-seq, where 10–35% of the reads map to rRNA (20). Second, given the stranded nature of the single cell total RNA sequencing data, quantification of antisense genes is accurate, which is not possible when using unstranded data. In contrast to the three existing single cell total RNA-seq methods, our method uniquely combines these two features that are highly desirable for total RNA-sequencing (19–21). Third, as expected, our single cell total RNA libraries contain substantially more intronic reads compared to polyA[+] RNA libraries (41,42). Such intronic reads can be used to detect changes in nascent transcription, whereby the difference in exonic and intronic reads provides insights in post-transcriptional regulation (43). As such, we believe that our method may be particularly well suited for 'RNA velocity analysis' of single cells (44). Fourth, the single cell total RNA-seq workflow presented in this paper detects relatively more protein coding genes, pseudogenes, lincRNAs and miscellaneous RNA (miscRNA) compared to single cell polyA[+] RNA libraries, when corrected for equal sequencing depth. While the number of detected genes increases with sequencing depth, there seems to be no plateau yet at 8 million reads, suggesting that further increasing the sequencing depth, could enable low abundant gene detection. Fifth, our method also detects non-polyadenylated

RNA molecules, such as histone genes, lncRNAs and circRNAs. In the NGP dataset, 537 circRNAs were detected using reads with evidence for back splicing. In order to detect more circRNAs in an individual cell, a higher sequencing depth is required or libraries should be enriched for circRNAs by selectively removing linear RNA by exonuclease treatment prior to library prep and sequencing (18). Sixth, the data enables reference guided transcriptome assembly, resulting in the detection of 5360 novel genes. Finally, differential gene expression analysis and gene set enrichment of NGP cells treated with nutlin-3 confirmed activation of the TP53 pathway at the transcriptional level.

One limitation of the implementation of the single cell total RNA library preparation method on the C1 instrument is the relatively low throughput, as maximally 96 cells are simultaneously captured. In contrast, current droplet-based single cell methods capture thousands of individual cells, but these systems are limited to 3′ end sequencing of polyadenylated RNA, preventing quantification of splice variants and non-polyadenylated transcripts. To enable the analysis of higher cell numbers, we demonstrated that the method works equally well on FACS sorted cells in microplates. By using FACS sorted cells the throughput can be increased and no specialized devices, such as the C1, are required. Finally, an advantage of our total RNA-seq protocol on both C1 and in microplates is that single-end sequencing is sufficient while more expensive paired-end sequencing is required for most droplet-based methods. We advice to use the single cell total RNA-seq method rather than polyA[+] methods if it is desired to study non-polyadenylated RNA molecules such as lncRNAs or circRNAs, if strand-specific data is a must and if full transcript sequencing is priority (e.g. analysis of alternative splicing, RNA editing or somatic mutations).

DATA AVAILABILITY

The SMARTer single cell total RNA sequencing script is deposited in Script Hub (Fluidigm). The fastq files and
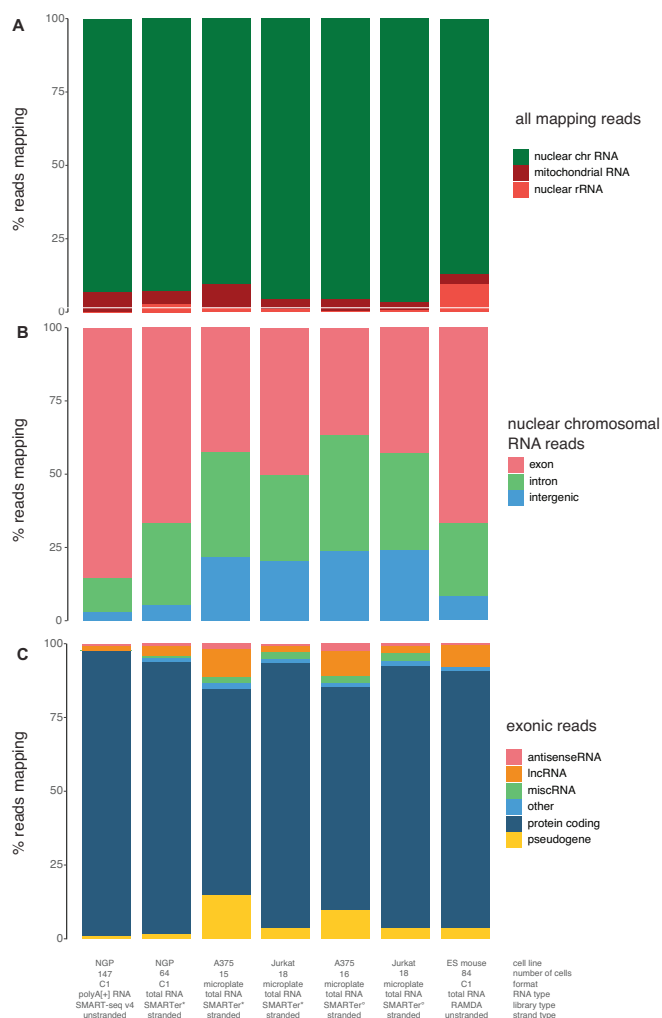
Figure 8. Mean read distributions are similar for total RNA sequencing libraries generated on C1 or in microplates. (**A**) Mean percentage of reads derived from nuclear RNA, mitochondrial RNA and ribosomal RNA quantified with STAR. Single cell total RNA libraries of the FACS sorted cells were generated using two different reagent kits (#634413, denoted with * and #634444, denoted with °). (**B**) Mean percentage of reads originating from nuclear chromosomes derived from exonic, intronic and intergenic regions quantified with STAR. (**C**) Mean percentage of exonic reads attributed to the different RNA biotypes quantified with Kallisto.

processed data is available through GEO (GSE119984 and GSE130578).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

FUNDING

REFERENCES

1. Tang,F., Barbacioru,C., Wang,Y., Nordman,E., Lee,C., Xu,N., Wang,X., Bodeau,J., Tuch,B.B., Siddiqui,A. *et al.* (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods*, **6**, 377–382.
2. Junker,J.P. and van Oudenaarden,A. (2014) Every cell is special: genome-wide studies add a new dimension to single-cell biology. *Cell*, **157**, 8–11.
3. Islam,S., Kjällquist,U., Moliner,A., Zajac,P., Fan,J.-B., Lönnerberg,P. and Linnarsson,S. (2011) Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.*, **21**, 1160–1167.
4. Soumillon,M., Cacchiarelli,D., Semrau,S., van Oudenaarden,A. and Mikkelsen,T.S. (2014) Characterization of directed differentiation by high-throughput single-cell RNA-Seq - SI. bioRxiv doi: 10.1101/003236, 05 March 2014, preprint: not peer reviewed.
5. Picelli,S., Faridani,O.R., Björklund,Å.K., Winberg,G., Sagasser,S. and Sandberg,R. (2014) Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.*, **9**, 171–181.
6. Picelli,S., Björklund,Å.K., Faridani,O.R., Sagasser,S., Winberg,G. and Sandberg,R. (2013) smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098.
7. Hashimshony,T., Wagner,F., Sher,N. and Yanai,I. (2012) CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.*, **2**, 666–673.
8. Hashimshony,T., Senderovich,N., Avital,G., Klochendler,A., de Leeuw,Y., Anavy,L., Gennert,D., Li,S., Livak,K.J., Rozenblatt-Rosen,O. *et al.* (2016) CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.*, **17**, 77.
9. Jaitin,D.A., Kenigsberg,E., Keren-Shaul,H., Elefant,N., Paul,F., Zaretsky,I., Mildner,A., Cohen,N., Jung,S., Tanay,A. *et al.* (2014) Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science*, **343**, 776–779.
10. Zheng,S., Papalexi,E., Butler,A., Stephenson,W. and Satija,R. (2018) Molecular transitions in early progenitors during human cord blood hematopoiesis. *Mol. Syst. Biol.*, **14**, e8041.
11. Rosenberg,A.B., Roco,C.M., Muscat,R.A., Kuchina,A., Sample,P., Yao,Z., Graybuck,L.T., Peeler,D.J., Mukherjee,S., Chen,W. *et al.* (2018) Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*, **360**, 176–182.
12. Macosko,E.Z., Basu,A., Satija,R., Nemesh,J., Shekhar,K., Goldman,M., Tirosh,I., Bialas,A.R., Kamitaki,N., Martersteck,E.M. *et al.* (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, **161**, 1202–1214.

13. Zilionis,R., Nainys,J., Veres,A., Savova,V., Zemmour,D., Klein,A.M. and Mazutis,L. (2016) Single-cell barcoding and sequencing using droplet microfluidics. *Nat. Protoc.*, **12**, 44–73.

14. Zheng,G.X.Y., Terry,J.M., Belgrader,P., Ryvkin,P., Bent,Z.W., Wilson,R., Ziraldo,S.B., Wheeler,T.D., McDermott,G.P., Zhu,J. *et al.* (2017) Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, **8**, 14049.

15. Kashima,Y., Suzuki,A., Liu,Y., Hosokawa,M., Matsunaga,H., Shirai,M., Arikawa,K., Sugano,S., Kohno,T., Takeyama,H. *et al.* (2018) Combinatory use of distinct single-cell RNA-seq analytical platforms reveals the heterogeneous transcriptome response. *Sci. Rep.*, **8**, 3482.

16. Yang,L., Duff,M.O., Graveley,B.R., Carmichael,G.G. and Chen,L.-L. (2011) Genomewide characterization of non-polyadenylated RNAs. *Genome Biol.*, **12**, R16.

17. Lai,F., Gardini,A., Zhang,A. and Shiekhattar,R. (2015) Integrator mediates the biogenesis of enhancer RNAs. *Nature*, **525**, 399–403.

18. Jeck,W.R. and Sharpless,N.E. (2014) Detecting and characterizing circular RNAs. *Nat. Biotechnol.*, **32**, 453–461.

19. Fan,X., Zhang,X., Wu,X., Guo,H., Hu,Y., Tang,F. and Huang,Y. (2011) Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos.*Genome Biol.*, **16**, 148.

20. Hayashi,T., Ozaki,H., Sasagawa,Y., Umeda,M., Danno,H. and Nikaido,I. (2018) Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. *Nat. Commun.*, **9**, 619.

21. Sheng,K., Cao,W., Niu,Y., Deng,Q. and Zong,C. (2017) Effective detection of variation in single-cell transcriptomes using MATQ-seq. *Nat. Methods*, **14**, 267–270.

22. Balbin,O.A., Malik,R., Dhanasekaran,S.M., Prensner,J.R., Cao,X., Wu,Y.-M., Robinson,D., Wang,R., Chen,G., Beer,D.G. *et al.* (2015) The landscape of antisense gene expression in human cancers. *Genome Res.*, **25**, 1068–1079.

23. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10.

24. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.

25. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

26. Zerbino,D.R., Achuthan,P., Akanni,W., Amode,M.R., Barrell,D., Bhai,J., Billis,K., Cummins,C., Gall,A., Girón,C.G. *et al.* (2018) Ensembl 2018. *Nucleic Acids Res.*, **46**, D754–D761.

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

28. Volders,P.J., Verheggen,K., Menschaert,G., Vandepoele,K., Martens,L., Vandesompele,J. and Mestdagh,P. (2015) An update on LNCipedia: a database for annotated human lncRNA sequences. *Nucleic Acids Res.*, **43**, 4363–4364.

29. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.

30. Van den Berge,K., Perraudeau,F., Soneson,C., Love,M.I., Risso,D., Vert,J.-P., Robinson,M.D., Dudoit,S. and Clement,L. (2018) Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome Biol.*, **19**, 24.

31. Memczak,S., Jens,M., Elefsinioti,A., Torti,F., Krueger,J., Rybak,A., Maier,L., Mackowiak,S.D., Gregersen,L.H., Munschauer,M. *et al.* (2013) Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature*, **495**, 333–338.

32. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.

33. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.

34. Pertea,M., Pertea,G.M., Antonescu,C.M., Chang,T.-C., Mendell,J.T. and Salzberg,S.L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.*, **33**, 290–295.

35. Buettner,F., Natarajan,K.N., Casale,F.P., Proserpio,V., Scialdone,A., Theis,F.J., Teichmann,S.A., Marioni,J.C. and Stegle,O. (2015) Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.*, **33**, 155–160.

36. Tung,P.-Y., Blischak,J.D., Hsiao,C.J., Knowles,D.A., Burnett,J.E., Pritchard,J.K. and Gilad,Y. (2017) Batch effects and the effective design of single-cell gene expression studies. *Sci. Rep.*, **7**, 39921.

37. Cabili,M.N., Trapnell,C., Goff,L., Koziol,M., Tazon-Vega,B., Regev,A. and Rinn,J.L. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.*, **25**, 1915–1927.

38. Livyatan,I., Harikumar,A., Nissim-Rafinia,M., Duttagupta,R., Gingeras,T.R. and Meshorer,E. (2013) Non-polyadenylated transcription in embryonic stem cells reveals novel non-coding RNA related to pluripotency and differentiation. *Nucleic Acids Res.*, **41**, 6300–6315.

39. Hsieh,C.L., Donlon,T.A., Darras,B.T., Chang,D.D., Topper,J.N., Clayton,D.A. and Francke,U. (1990) The gene for the RNA component of the mitochondrial RNA-processing endoribonuclease is located on human chromosome 9p and on mouse chromosome 4. *Genomics*, **6**, 540–544.

40. Fish,R.N., Bostick,M., Lehman,A. and Farmer,A. (2016) Transcriptome analysis at the single-cell level using smart technology. In: *Current Protocols in Molecular Biology*. John Wiley & Sons, Inc., Hoboken, Vol. **116**, pp. 4.26.1–4.26.24.

41. Ameur,A., Zaghlool,A., Halvardson,J., Wetterbom,A., Gyllensten,U., Cavelier,L. and Feuk,L. (2011) Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nat. Struct. Mol. Biol.*, **18**, 1435–1440.

42. Zhao,S., Zhang,Y., Gamini,R., Zhang,B. and von Schack,D. (2018) Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. *Sci. Rep.*, **8**, 4781.

43. Gaidatzis,D., Burger,L., Florescu,M. and Stadler,M.B. (2015) Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation. *Nat. Biotechnol.*, **33**, 722–729.

44. La Manno,G., Soldatov,R., Zeisel,A., Braun,E., Hochgerner,H., Lidschreiber,K., Kastriti,M.E., Lönnerberg,P., Furlan,A. *et al.* (2018) RNA velocity of single cells. *Nature*, **560**, 494–498.

45. Macosko,E.Z., Basu,A., Satija,R., Nemesh,J., Shekhar,K., Goldman,M., Tirosh,I., Bialas,A.R., Kamitaki,N., Martersteck,E.M. *et al.* (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, **161**, 1202–1214.

46. Klein,A.M., Mazutis,L., Akartuna,I., Tallapragada,N., Veres,A., Li,V., Peshkin,L., Weitz,D.A. and Kirschner,M.W. (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, **161**, 1187–1201.

47. Sasagawa,Y., Nikaido,I., Hayashi,T., Danno,H., Uno,K.D., Imai,T. and Ueda,H.R. (2013) Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biol.*, **14**, 3097.

48. Fan,H.C., Fu,G.K. and Fodor,S.P.A. (2015) Combinatorial labeling of single cells for gene expression cytometry. *Science*, **347**, 1258367.