Performance of the cellenCHIP 384-3'RNA-Seq Kit

Johannes W. Bagnoli, PhD
Raphael Schneider, MSc
François Monjaret, PhD
Cellenion SASU, 60 Avenue Rockefeller, 69008 Lyon, France

Abstract

Over the last decade single cell RNA sequencing (scRNA-seq) has become a wide spread tool to analyze global expression levels of single cells with more and more protocols published each year. Cellenion’s cellenCHIP 384-3’RNA-Seq Kit is designed to create high quality single cell 3’transcriptome libraries using the cellenONE® technology and the cellenCHIP 384 nanowell substrate. In this study, the key performance characteristics of the cellenCHIP 384-3’RNA-Seq Kit are presented and compared to established competitive Kits as well as non-commercial methods. For a multi-parameter comparison and their combined picture, the performance score was developed as a benchmark tool. The performance score revealed a superior performance of the cellenCHIP 384-3’RNA-Seq Kit when compared to competitor methods. In addition, the cellenCHIP 384 in combination with the cellenONE® single cell isolation and dispensing technology reliably inhibits cell to cell crosstalk and reduces background noise.
Introduction

Every cell in a multicellular organism expresses a unique subset of the genes found in the genome, enabling a diverse and complex landscape of various cell types and functions. Expressed genes are transcribed into messenger RNA molecules, which are further translated to build specific proteins. The gene expression profile of each cell and hence the messenger RNA content varies depending on cell state and function. The 3’RNA sequencing method focuses on the abundance and relative changes in gene expression across sample groups. Applied to individual cells, it offers an insight into their molecular condition, thereby revealing their current state and function. However, there are multiple challenges to access this level of information, with 5 major bottlenecks:

1. Access to the single cell level
2. Tracking the cell origin of each RNA transcript
3. Collecting and creating enough cell material to fit the minimum sequencing input
4. Reducing the impact of the protocol on the transcript level (technical noise)
5. Keeping the price per cell low

Over the last decade, several protocols and technical solutions for scRNA-seq have been published (Mereu et al., 2020; Ziegenhain et al., 2017) or put on the market.

Various approaches have been developed to find the best solution to clear the 5 bottlenecks cited above. The competitor methods (referred here as Comp.1-4) used in this study are summarized below and in table 1:

a) **Comp1 and Comp.2**: The most used technology today is based on microfluidics and microdroplets. The sample (a cell suspension) is partitioned in droplets and each droplet is considered as a virtual well where RNAs are converted in cDNA and tagged, before being pooled, amplified using PCR and sequenced. Most of the 5 bottlenecks are addressed, however the access to the single cell level is based only on distribution statistics, resulting in an uncertainty about the quality of the data. Another insecurity of the method results from the lack of a process control: It’s not possible to monitor what is happening to the sample.

b) **Comp.3**: Another player in the scRNA-seq field developed a miniaturized version of Smart-seq2, the gold standard protocol for full length RNA-seq (Picelli et al., 2014) Despite the advantage of being based on this robust protocol, cell isolation is based on Poisson law, reducing the value of the produced data. Also, the price per cell is higher than for other methods.

c) **Comp.4**: A third protocol is based on *in vitro* transcription (IVT) to amplify the material to sequence instead of PCR based amplification. This reduces an inherent bias introduced by PCR: exponential amplification of synthesis errors. It is based on the CEL-seq2 protocol (Hashimshony et al., 2016) and is performed in plates. Main limitations are the technology used for isolating cells, and price per cell.

d) **Smart-seq2 and Smart-seq3** (Hagemann-Jensen et al., 2020) are plate-based solutions which do not require any specific instrumentation and have been incorporated into this study. They are both based on PCR and do not include any solution to isolate cells but are currently considered as the most powerful.
Materials and methods

Design of experiments
Single human and mouse cells (HEK293T and NIH3T3) were isolated using the cellenONE® single cell isolation and dispensing device into a cellenCHIP 384 (Cellenion, CEC-5015-4) included in the cellenCHIP 384-3’RNA-Seq Kit (Cellenion, CTR-5016-1-4; 5-8; 9-12; 13-16, see Figure 1) using the cellenCHIP 384 Accessory Kit (Cellenion, CAK-5015). A checkerboard pattern of human and mouse cells was used. Moreover, for each set of 96 wells, 3 positive controls for each cell type were prepared by dispensing exactly 5 cells per well. Four different negative controls (3 wells per control) were prepared as following: (i) dispensing a single human cell without RT enzyme, (ii) no cell dispensing, (iii) dispensing culture medium without any cell, and (iv) dispensing a single human cell with RT buffer that contained RNase.

Cell culture conditions
HEK293T and NIH3T3 cells were cultured in high-glucose Dulbecco’s modified Eagle’s medium (Thermo Fisher) supplemented with 10 % fetal bovine serum (FBS, DUTSCHER), 100 U/ml penicillin (Corning), 100 μg/ml streptomycin (Corning) and 250 μg/ml amphotericin B (Corning). Passaging was performed routinely every 3 days using Trypsin-EDTA (0.05%, Thermo Fisher).

Isolation of cells using the cellenONE® technology
The cellenONE® is an automated single cell isolation and dispensing device that has been used for the isolation of single cells in the cellenCHIP 384-3’RNA-Seq Kit protocol. In brief, cells were resuspended in PBS just prior isolation. After loading in the Piezo Dispensing Capillary (PDC), the software was set to detect all particles between 10 and 100 μm in size, a maximum elongation factor of 4 and a maximum circularity factor of 3. To isolate single cells, particles with a size between 19 to 27 μm in diameter, an elongation factor below 2 and a circularity factor below 1.1 were selected. The reverse factor was set to control for free floating RNA and DNA by dispensing of isolation media without cells. Cells were dispensed using the cellenONE® in a checkerboard pattern in the following sequence: 1) Multiple NIH3T3 cells (5 per well), 2) single NIH3T3 cells, 3) Media from HEK293T cells, 4) multiple HEK293T cells (5 per well), 5) single HEK293T cells.

cellenCHIP 384-3’RNA-Seq Kit and library generation
The cellenCHIP 384-3’RNA-Seq Kit is a complete solution for single cell library preparation. It includes the cellenCHIP 384, consisting of 4 identical 96 well arrays that contain in each nanowell oligo-dt-primer, an individual cell barcode (CB) and a unique molecular identifier (UMI) for cDNA
generation. With the barcodes and the UMIs the protocol allows the identification of every individual cell transcript. The sample preparation and the cDNA generation take place in the cellenCHIP 384. First, the barcode oligos are rehydrated with the Lysis & RT Buffers. Single cells were then isolated and dispensed into the cellenCHIP 384 using the cellenONE®.

After cell isolation, wells of the cellenCHIP 384 were sealed, and reverse transcription was performed on an in-situ block at 42°C for 90 minutes. Afterwards, the reverse transcription products of each well were collected and pooled by inverting the cellenCHIP 384 and centrifugation into a recovery funnel prior to transfer to microcentrifuge tubes. The cDNA was quantified using the Qubit™ 1X dsDNA HS Assay Kit (Thermo Fisher) in combination with the Qubit 4 Fluorometer (Thermo Fisher). The cDNA was amplified for a maximum of 18 PCR cycles, and amplified cDNA was used to generate Illumina sequencing libraries with one-sided tagmentation and PCR amplification. The libraries were then analyzed for size distributions on high-sensitivity D5000 DNA chips (Agilent TapeStation). Libraries were sequenced using pair-end 20-8-50 (NextSeq 500, Illumina).

Sequencing
Libraries were paired end sequenced on a high output flow cell of an Illumina NextSeq 500 instrument. Cellular and molecular barcodes were obtained by with the first read (20 nt) while the second read (50 nt) obtained the sequence information of the cDNA fragment. An additional 8 base i7 barcode read was done for demultiplexing of the sequencing lane.

Data of competitors
Different datasets from established competitive 3'RNA-Seq Kits as well as non-commercial methods were downloaded from publicly available databases on HEK cells.

Table 1: Summary of competitor methods used in this study

<table>
<thead>
<tr>
<th>Name in this AppNote</th>
<th>Cell isolation</th>
<th>Library preparation</th>
<th>Chemistry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comp.1</td>
<td>Poisson law using microdroplets in microfluidic</td>
<td>In microdroplets</td>
<td>3'RNA-seq based on UMI and PCR, second version</td>
</tr>
<tr>
<td>Comp.2</td>
<td>Poisson law using microdroplets in microfluidic</td>
<td>In microdroplets</td>
<td>3'RNA-seq based on UMI and PCR, second version</td>
</tr>
<tr>
<td>Comp.3</td>
<td>Poisson law using nano-dispensing</td>
<td>In miniaturized plate</td>
<td>3'RNA-seq based on UMI and PCR</td>
</tr>
<tr>
<td>Comp.4</td>
<td>Not included</td>
<td>In regular plate</td>
<td>3'RNA-seq based on UMI and IVT</td>
</tr>
<tr>
<td>Smart-seq2</td>
<td>Not included</td>
<td>In regular plate</td>
<td>Full-length RNA-seq based on PCR</td>
</tr>
<tr>
<td>Smart-seq3</td>
<td>Not included</td>
<td>In regular plate</td>
<td>Full-length RNA-seq based on PCR plus 5'RNAseq based on UMI</td>
</tr>
</tbody>
</table>

Primary Data Analysis
All raw fastq data was processed using zUMIs v2.9.3e (Parekh et al., 2018) in combination with STAR v2.7.3a (Dobin et al., 2013). Reads were mapped to the human (hg38) or mouse (mm10)
reference genome and gene annotations were obtained from gencode.v34 (GRCh38) and gencode.vM2 (GRCm38.p6) respectively. Down sampling to fixed numbers of raw sequencing reads per cell was performed using the “-d” option in zUMIs.

**Performance score calculation**

![Diagram of performance score calculation](image)

RNA sequencing data is highly complex and needs to be carefully analyzed when looking at the efficiency of different methods. In this study, the following seven different parameters (Figure 2) were systematically selected that have been used in all methods to validate the quality of the dataset: the Mapping Rate (MR), the fraction exon/intron (Fei), the unique reads (Ur), the Gene per read (Gpr), the proportion of rRNA reads (rRNA), the Percent deviation (Pdev) and the Dropout (Drop). Calculation methods are described in Table 2. (MR), (Fei), (Ur), (Gpr) and (rRNA) are calculated for each single cell (i), while (Pdev) and (Drop) are calculated for each library (j). The 7 parameters are combined with an equal weight to define the performance score ranking from 0 (very poor) to 100 (very good). A Performance score was calculated for each method on strictly comparable samples down sampled to 20 K reads/cell.

**Table 2: Calculation of the performance score**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mapping Rate (MR):</td>
<td>( MR_i = \frac{\text{Exonic Reads}_i + \text{Intronic Reads}_i}{\text{Total Reads}_i} )</td>
</tr>
<tr>
<td>Fraction exon/intron (Fei):</td>
<td>( Fei_i = \frac{\text{Exonic Reads}_i}{MR_i} )</td>
</tr>
<tr>
<td>Unique reads (Ur):</td>
<td>( Ur_i = \frac{\text{Unique Reads}_i}{\text{Total Reads}_i} )</td>
</tr>
<tr>
<td>Proportion rRNA reads (rRNA):</td>
<td>( rRNA_i = \frac{\text{rRNA Reads}_i}{\text{Total Reads}_i} )</td>
</tr>
<tr>
<td>Gene per read (Gpr):</td>
<td>( Gpr_i = \frac{\text{Number of Genes}_i}{\text{Total Reads}_i} )</td>
</tr>
<tr>
<td>Percent deviation (Pdev):</td>
<td>( Pdev_i = \frac{\text{std}}{\text{number of genes}_i} )</td>
</tr>
<tr>
<td>Dropout (Drop):</td>
<td>( Drop_i = \frac{\text{UMI}<em>{j=0}}{\text{UMI}</em>{j&gt;0}} )</td>
</tr>
</tbody>
</table>

**Performance Score**

\[
Performance\ Score_{ij} = \frac{MR_i + Fei_i + Ur_i + Gpr_i + rRNA_i + (1 - Pdev_i) + (1 - Drop_j)}{7} \times 100
\]

**Cross contamination**

Species contamination was analyzed by mapping of raw reads to several genomes including human mouse and rat using FastQ-Screen-0.14.1 with default parameters.
Results and discussion

The cellenCHIP 384-3’RNA-Seq Kit produces high quality data

First, the key performance characteristics of the cellenCHIP 384-3’RNA-Seq Kit were analysed (Figure 3-5) and then compared to established competitive 3’RNA-seq Kits as well as non-commercial methods (Figure 6).

The cellenCHIP 384 nanotechnology efficiently minimizes cell to cell crosstalk

To investigate the quality and robustness of the cellenCHIP 384-3’RNA-Seq Kit using both the cellenCHIP 384 nanowell substrate and the cellenONE® nanoliter and single cell dispenser, a checkerboard pattern of human and mouse cells was elaborated.

As the cellenONE® and the cellenCHIP 384 allow to perfectly control which cell is dispensed in each well, crosstalk would be detected if human mRNAs were barcoded with mouse assigned barcodes, and vice versa.

For each well, the uniquely aligning reads to the human and mouse genome were calculated to determine if any contamination between wells occurred. For all wells, the vast majority (>94 %, triplicate) of reads only mapped to the corresponding genome with human transcripts found in wells containing human cells and vice versa for wells containing mouse cells (Figure 3). Only 4 cells diverge from the expected species in samples that comprised a low overall number of reads. These observations confirm that no detectable cross contamination between cellenCHIP 384 wells or within the cellenONE® dispensing nozzle occurred, demonstrating the very low crosstalk using this protocol.

Figure 3: Cross contamination between wells
Schematic representation of the human/mice cell checkerboard pattern. For each cell barcode, uniquely aligning reads to human or mouse gene features are shown in a dot plot. Square, disc and triangle point shapes represent three different technical replicates. Only 4 cells across 234 (1.7%) diverge from planed species. Blue points represent correctly assigned human cells, pink points represent correctly assigned mouse cells. Grey points represent contaminated cells or cells with a sequencing depth below the analysis threshold (20K reads).
The cellenONE® technology allows flexible and controlled dispensing

The cellenONE® dispensing technology is very flexible and allows to include positive controls and several negative controls in a single cellenCHIP 384. The positive controls consist of 5 cells dispensed in a single well, increasing the total amount of RNA in the well thus ensuring to have an efficient reaction in the chip. To control for any kind of contamination from the RT mix, a non-template control (No Cell) was included. To check whether free RNA within the cell solution might lead to an increase of cross contamination drops of the media without cells were dispensed (Media). In addition, RNase was added into the (Media) well destroying the targeted substrate to test whether fragments from genomic DNA were obtained. As negative controls, wells which did not have any reverse transcriptase enzyme in the mix were included (No RT).

As expected, more reads were detected for all positive controls (Figure 4) that contained 5 instead of only one cell and thus lead to the production of more cDNA per well, in a robust and repeatable manner. On the contrary, very few reads were detected across all four negative controls (Figure 4) and their number were consistent within the four different conditions. Accordingly, those negative controls confirmed the absence of contamination between wells and within buffers, the absence of amplification from gDNA and the absence of free-floating mRNA in the cell suspension. This last (No RT) control is highly valuable as it allows to evaluate and potentially remove background noise coming from free floating mRNA in the cell suspension for samples that cannot be cleaned or washed.

A high number of genes per cells can be detected using the cellenCHIP 384-3’RNA-Seq Kit

The number of genes detected per single cell has been widely used as the major measurement parameter for assigning the sensitivity of any scRNA-seq method (Mereu et al., 2020; Ziegenhain et al., 2017). Hence, the number of detected genes (UMI>=1) per HEK cells was measured using the cellenCHIP 384-3’RNA-Seq Kit over various sequencing depths using down sampling (Figure 4). It could be shown, that even at low sequencing depths, the Kit reliably detects a high number of genes per cell, with a minimum of 1895 genes detected at 5000 raw sequencing reads and a maximum of 7409 genes detected at the highest sequencing depth here of 200K raw reads. This indicates the high sensitivity of the method even at low sequencing depths.

Although this parameter is commonly used as the main sensitivity marker, its use is not without complications. For example, the number of detected genes could be artificially inflated by genomic priming, PCR chimeras, leftover oligo-dT-primers during amplification or cell to cell contamination (Dixit, 2016; Kalle et al., 2014; Ziegenhain et al., 2021). Therefore, the quality of the libraries was further investigated using various quality measurements to verify the accuracy of the cellenCHIP 384-3’RNA-Seq Kit.

Figure 4: Number of detected genes
Number of detected Genes (exon and intron mapped reads) in libraries generated from 117 single HEK cells using the cellenCHIP 384-3’RNA-Seq Kit when down sampled to different numbers of raw sequence reads. Each box represents the median and first and third quartiles per cell per sequencing depth. Whiskers indicate the most extreme data point that is no more than 1.5 times the length of the box away from the box.
Low intron contamination can be found using the cellenCHIP 384-3’RNA-Seq Kit

Data from 3’scRNA-sequencing should be comprised largely of exonic mapping reads (Lee et al., 2020), as full matured mRNA transcripts generally lack any other genomic feature, such as introns and non-coding elements. Hence, the data was analyzed concerning three major parameters: i) The percentage of reads mapping to either exonic or intronic regions (Mapping Rate) ii) The fraction of exonic reads within the mapped reads (Fraction Exon/Intron) and iii) the percentage of reads which map to a unique genomic locus (Unique Reads). For all three Parameters the cellenCHIP 384-3’RNA-Seq Kit showed high quality scores with ~80% of reads mapping to exonic and intronic regions and >70% within these to exonic regions (Figure 5). In addition, >70% of reads mapped to a single genomic locus (Figure 5). This is in accordance with parameters of the best solutions available so far. Taken together this shows that the method captures mRNA molecules very efficiently and does not generate sporadic fragments from genomic DNA.

![Figure 5: Quality of scRNA-seq libraries](image)

The cellenCHIP 384-3’RNA-Seq Kit performs robustly across wells

In a homogeneous cell population, cells mostly express the same genes at a similar rate. While there are cell specific expression patterns, such as cell cycle regulatory genes, the same number of genes should be detected in the majority of cells. To investigate this, the variability in the number of detected genes across the cells was analyzed. In addition, it was analyzed how often each gene was not detected over all cells (Dropout). While the dropout is commonly very high in scRNA-sequencing data sets (zero inflation) (Ziegenhain et al., 2018) a relatively low high of was observed for the library with ~83 % (Figure 5), which is in the same range as the best solutions on the market. Combining the observed low variability between cells with the unique number of detected genes (Figure 5), it can be concluded that the cellenCHIP 384-3’RNA-Seq Kit performs very robustly and does not show a high cell to cell variance, leading to a decreased noise and a better power to identify differences between groups of cells.
The cellenCHIP 384-3´RNA-Seq Kit outperforms competitor methods

To show the power and the advantages of the cellenCHIP 384-3´RNA-Seq Kit, the results were compared to public available data from HEK cells analyzed with four different competitors methods of 3’RNA-Seq library preparation protocols, in microfluidic and plate based format (Mereu et al., 2020). To have a full understanding and clear picture of the quality of these methods the previously mentioned parameters Mapping Rate (MR), the fraction exon/intron (Fei), the unique reads (Ur), the Gene per read (Gpr), the proportion of rRNA reads (rRNA), the Percent deviation (Pdev) and the Dropout (Drop) were analyzed for all competitor methods at a sequencing depth of 20K reads per cell. These parameters were compared to a final performance score (see methods).

Overall, the cellenCHIP 384-3´RNA-Seq Kit showed a superior performance when compared to 3´scRNA-Seq solutions of competitors (Figure 6 Comp.1- Comp.4). It outperformed all four of them, showing a significantly higher mean performance score ($p = 2.3 \times 10^{-10}$, $7.2 \times 10^{-12}$, $2.2 \times 10^{-16}$, $2.2 \times 10^{-16}$, Welch two sample t-test). In addition, when compared to the gold standard version of full-length RNA-seq, Smart-seq 2, and its revised version, Smart-seq 3 the cellenCHIP 384-3´RNA-Seq Kit outperformed the current gold standard ($p = 0.004$, Welch two sample t-test). Only Smart-Seq 3 showed a higher performance score than with the cellenCHIP 384-3´RNA-Seq Kit solution ($p = 2.2 \times 10^{-16}$, Welch two sample t-test). As Smart-seq 3 is the most sensitive method to date, it should be considered as maximum achievable performance score and used as the target any solution should aim for.

![Figure 6: Performance Score measurement of scRNA-seq Kits and methods](image)

Performance Score values of libraries generated from HEK cells at 20K reads/cell using several commercially available 3´scRNA-seq Kits, including the cellenCHIP 384-3´RNA-Seq Kit, or the gold standard full length scRNA-seq method and its revised version, Smart-seq 2 and Smart-seq 3. Each box represents the median and first and third quartiles per cell per sequencing depth. Whiskers indicate the most extreme data point that is no more than 1.5 times the length of the box away from the box.
Conclusion and future direction

The cellenCHIP 384-3’RNA-Seq Kit combines the cellenONE® technology for single cell isolation and dispensing as well as the cellenCHIP 384 nanowell substrate to enable miniaturization of single cell 3’transcriptome in nanoliter volumes to generate data with high sensitivity and low background. In this study, the high performance of the cellenCHIP 384-3’RNA-Seq Kit could be shown in comparison to other automated 3’RNA-Seq methods as well as full length RNA-seq library preparation protocols.

➢ The cellenCHIP 384-3’RNA-Seq Kit including the unique design of the cellenCHIP 384 prevents well-to-well contamination
➢ Single cell isolation and dispensing using the cellenONE® technology enables flexibility to include additional controls for efficient and clean sorting of single cells without background artefacts
➢ The cellenCHIP 384-3’RNA-Seq Kit reliably produces highly sensitive libraries even at low sequencing depth
➢ The cellenCHIP 384-3’RNA-Seq Kit shows high specificity in capturing mRNA transcripts and does not suffer from molecular artefacts, reducing noise
➢ When compared to competitor methods the cellenCHIP 384-3’RNA-Seq Kit shows superior performance
References


