CLEAR: coverage-based limiting-cell experiment analysis for RNA-seq

Logan A. Walker[1,2], Michael G. Sovic[2], Chi-Ling Chiang[2,3], Eileen Hu[2,3], Jiyeon K. Denninger[4], Xi Chen[2], Elizabeth D. Kirby[4,5], John C. Byrd[2,3], Natarajan Muthusamy[2,3], Ralf Bundschuh[1,3,6,7*] and Pearlly Yan[2,3*]

Abstract

Background: Direct cDNA preamplification protocols developed for single-cell RNA-seq have enabled transcriptome profiling of precious clinical samples and rare cell populations without the need for sample pooling or RNA extraction. We term the use of single-cell chemistries for sequencing low numbers of cells limiting-cell RNA-seq (lcRNA-seq). Currently, there is no customized algorithm to select robust/low-noise transcripts from lcRNA-seq data for between-group comparisons.

Methods: Herein, we present CLEAR, a workflow that identifies reliably quantifiable transcripts in lcRNA-seq data for differentially expressed genes (DEG) analysis. Total RNA obtained from primary chronic lymphocytic leukemia (CLL) CD5+ and CD5− cells were used to develop the CLEAR algorithm. Once established, the performance of CLEAR was evaluated with FACS-sorted cells enriched from mouse Dentate Gyrus (DG).

Results: When using CLEAR transcripts vs. using all transcripts in CLL samples, downstream analyses revealed a higher proportion of shared transcripts across three input amounts and improved principal component analysis (PCA) separation of the two cell types. In mouse DG samples, CLEAR identifies noisy transcripts and their removal improves PCA separation of the anticipated cell populations. In addition, CLEAR was applied to two publicly-available datasets to demonstrate its utility in lcRNA-seq data from other institutions. If imputation is applied to limit the effect of missing data points, CLEAR can also be used in large clinical trials and in single cell studies.

Conclusions: lcRNA-seq coupled with CLEAR is widely used in our institution for profiling immune cells (circulating or tissue-infiltrating) for its transcript preservation characteristics. CLEAR fills an important niche in pre-processing lcRNA-seq data to facilitate transcriptome profiling and DEG analysis. We demonstrate the utility of CLEAR in analyzing rare cell populations in clinical samples and in murine neural DG region without sample pooling.

Keywords: Rare cells, Ultralow input, Pre-filtering, RNA-seq, Differential gene expression analysis

Background

Deep sequencing of transcriptomes (RNA-seq) provides important insights into biology and disease. Bulk RNA-seq requires hundreds of thousands of cells. The resultant transcriptomic profile therefore represents the average of cells at different transcriptomic states or even different cell types within the same tissues (e.g., infiltrating immune cells or normal cells in tumor samples). With the discovery of new genes and splice junctions in the first single-cell RNA-seq (scRNA-seq) study [1], researchers realized the need to profile single-cell transcriptomes. Coinciding with this intense interest is the development of diverse approaches to perform scRNA-seq, which have been summarized in recent reviews [2–5]. Improvements in reagents that enable full-length transcriptome

*Correspondence: bundschuh@mps.ohio-state.edu; Pearlly.Yan@osumc.edu
[1] Department of Physics, College of Arts and Sciences, The Ohio State University, Columbus, OH, USA
[3] Division of Hematology, Department of Internal Medicine, College of Medicine, The Ohio State University, Columbus, OH, USA
Full list of author information is available at the end of the article

Walker *et al. J Transl Med*   (2020) 18:63

Page 2 of 15

profiling by direct global amplification at the single-cell level (e.g., SMART-seq [6–8], Quartz-Seq [9], and the 'Tang' method [1]) have also enabled direct amplification for the analysis of groups of 10's–100's of cells, i.e., limiting-cell RNA-seq (lcRNA-seq).

The advantages of direct amplification are manifold. First, it lowers the barrier in identifying differentially expressed genes (DEGs) in rare cell populations. Until recently, researchers had to pool cells from multiple samples to extract quantifiable amounts of total RNA for library generation. With direct cDNA preamplification, the number of cells required for successful library preparation dropped below 100 cells, a level often achievable from just one sample. Also, quantifying RNA accurately below 250 pg/μL is challenging. Direct cDNA preamplification eliminates this need by using cell counts (e.g., from fluorescence activated cell sorting (FACS) or laser capture) instead of RNA mass to standardize input amounts between experimental groups. Second, direct cDNA preamplification greatly preserves transcript quality by quickly transforming labile RNA into stable cDNA, as degradation associated with extraction can be significant. Third, direct amplification of enriched cells deposited into well-plates allows the incorporation of nanoliter microfluidics to deliver/mix reagents and templates quickly. This further preserves RNA integrity.

Even with these advances, systems noise associated with transcript degradation is inevitable and requires computational solutions, especially if large numbers of replicates are not feasible. Current publications on lcRNA-seq data fall into two categories: cell-pool samples as part of method development or scRNA-seq methodology comparisons [6, 10–12] and ultralow amounts of extracted RNA as input for RNA-seq library preparation [13–15]. Analysis workflows for bulk RNA-seq and scRNA-seq data are distinct as each approach addresses a different research question. The goal of bulk RNA-seq is to identify differences in transcriptomic profiles between treatment groups, whereas the goal of scRNA-seq is to characterize cell subpopulations in tissues or bulk cells. In this regard, the aim of lcRNA-seq experiments is like that of bulk RNA-seq experiments, whereas their data quality (e.g., prevalence of zero count genes or 'dropout rate' [10, 16, 17]) is similar to that of scRNA-seq. Therefore, statistical methods often used for between-group comparisons in bulk RNA-seq studies, such as the negative binomial distribution-based test in DESeq2 [18], should not be used for lcRNA-seq data without modifications because they are susceptible to zero-count artifacts. On the other hand, the myriad of tools [19–30] for analyzing scRNA-seq data are tuned to work with high variabilities (e.g., true biological variations in single cells or variabilities due to cDNA preamplification) but require

large numbers of replicates not possible in lcRNA-seq studies.

Here, we describe CLEAR, a computational preprocessing approach for between-group comparisons of lcRNA-seq experiments. Designed for low numbers of replicates, CLEAR focuses on identifying robust transcripts with even read coverage for downstream analysis to control for data noise. It examines the noise pattern of individual samples to identify 'reliably quantifiable' transcripts for maximal signal and minimal noise. CLEAR transcripts common to replicates across two comparison groups will be used for subsequent analyses. Using a dataset derived from the same RNA stock but at dilutions spanning typical lcRNA-seq inputs, we show that CLEAR greatly improves similarity between results from three input RNA levels. In two public datasets, we demonstrate that the numbers and dispersion patterns of CLEAR transcripts yield a novel way to evaluate library qualities. In an in-house murine neural cell lcRNA-seq study, utilizing CLEAR transcripts significantly improves cell type separations by principal component analysis (PCA) and validations of cell phenotype markers. These case studies demonstrate the value of lcRNA-seq coupled with CLEAR in transcriptomic profiling of tissues found in rare niches and of precious clinical samples.

Methods
CLL patient sample acquisition
A chronic lymphocytic leukemia (CLL) patient sample was obtained from the Leukemia Tissue Bank (LTB), a shared resource of the NCI-funded OSU Comprehensive Cancer Center. The sample was obtained following written informed consent in accordance with the Declaration of Helsinki and under a protocol reviewed and approved by the Institutional Review Board of the Ohio State University. The patient had CLL as defined by the IWCLL 2008 criteria. The patient's white blood cells were isolated by Ficoll density gradient centrifugation (Ficoll-Paque Plus, Amersham Biosciences, Little Chalfont, UK) and samples were banked at -180 °C in liquid nitrogen. Frozen cells were thawed and washed with RPMI 1640 media (Gibco, Life Technologies, Grand Island, NY, USA) and resuspended at 5×10^6 cells/mL in complete medium containing 10% fetal bovine serum (FBS) (Sigma, St Louis, MO, USA), 2 mM L-glutamine, penicillin (100 U/mL), and streptomycin (100 μg/mL) (Gibco).

Animals for neural cell type analysis
All procedures involving animals were approved by the Ohio State Institutional Animal Care and Use Committee in accordance with institutional and national guidelines. Nestin-GFP mice were provided by Grigori Enikolopov at Cold Spring Harbor Laboratory [31]. All

mice were housed in a 12 h light–dark cycle with food and water ad libitum. For isolation of the dentate gyrus (DG), adult mice (6–9 week old) were anesthetized with an intraperitoneal injection of ketamine (87.5 mg/kg) and xylazine (12.5 mg/kg) before perfusion with PBS. Following perfusion, the brain was removed and placed in cold Neurobasal A medium (Gibco 10-888-022) on ice. After bisecting the brain along the midsagittal line, the cerebellum and diencephalic structures were removed to expose the hippocampus. Under a dissection microscope (Zeiss), the DG was excised using a beveled syringe needle and placed in ice cold PBS without calcium or magnesium (Gibco 10-010-049). DGs from mice were first mechanically dissociated with sterile scalpel blades and then enzymatically dissociated with a pre-warmed papain (Roche 10108014001)/dispase (Stem Cell Technologies 07913)/DNase (Stem Cell Technologies NC9007308) (PDD) cocktail at 37 °C for 20 min. Afterwards, the tissue was again mechanically disrupted by trituration for 1 min. Dissociated cells were collected by centrifugation at 500 g for 5 min.

Fluorescence activated cell sorting (FACS)

For the human sample all cells were stained and sorted by FACS Aria (BD Biosciences, San Jose, CA, USA). Live CLL B cells (CD5+ CD19+) and normal B cells (CD5− CD19+) were sorted from the patient sample. Briefly, cells for FACS were resuspended in PBS without calcium/magnesium and filtered through a 35 μm nylon filter and then stained for PE-conjugated CD45 (HI30), PerCp Cy-5.5-conjugated CD19 (HIB19), and Alexa Fluor 700-conjugated CD3 (UCHT1) monoclonal antibodies. Nonviable cells were excluded by the LIVE/DEAD Fixable Near-IR Dead Cell Stain Kit (Life Technologies, Carlsbad, CA, USA). Appropriate fluorescence minus one controls were used to determine nonspecific background staining. Single-cell gates were used to exclude the possibility of doublet cells. The FACS parameter diagrams for this process are available in Additional file 1: Figure S1.

For the Nestin-GFP mouse samples, cells were resuspended in PBS without calcium/magnesium and filtered through a 35 μm nylon filter before staining with the following antibodies: O4-APC (1:100, R&D FAB1326A), O1-eFluor660 (1:100, Thermo Fisher/eBioscience, Pittsburgh, PA, 5065082), GLAST-PE (1:50, Miltenyi Biotec, Bergisch Gladbach, Germany, 130-098-804), CD45-APC (1:100, BD Biosciences, Franklin Lakes, NJ, 561018), CD31-APC (1:100, BD Biosciences, Franklin Lakes, NJ, 561814). Cells were incubated with antibodies on ice for 30 min. During the last 10 min of staining, Hoechst dye (1:10,000, Thermo Fisher, Pittsburgh, PA, 33342) was added for live/dead discrimination. All cells were washed twice following staining and immediately sorted as stem,

progenitor, or astrocyte populations based on fluorescent markers with the FACSAria III (BD Biosciences, Franklin Lakes, NJ). CD31, CD45, O1, and O4 negative live cells were designated as stem cells if double positive for GLAST and GFP, progenitors if only GFP positive, and astrocytes if only GLAST positive. For cells in the limited cell number study, 50 cells were sorted into 96 well format plates for downstream lcRNA-seq library generation.

Total RNA extraction

Total RNA extraction was performed using Trizol reagent (Invitrogen, Carlsbad, CA). Briefly, approximately two million FACS-derived CD5+ and CD5− cells were separately sorted into 1.7 mL microcentrifuge tubes. Excess buffer was removed by centrifugation. Trizol reagent was added to cell pellets and the extraction protocol recommended by Invitrogen was followed. Total RNA was precipitated with 10 μg glycogen (Qiagen, Hilden, Germany). The quality of the total RNA was assessed with the Agilent 2100 Bioanalyzer (Agilent, Inc., Santa Clara, CA) using total RNA Pico chip.

Library generation and sample sequencing

Total CD5+ and CD5− RNA quantified using the Invitrogen Qubit RNA HS Assay kit (Invitrogen, Carlsbad, CA) was serially diluted to masses characteristic of single- and limiting-cell RNA-Seq (10-, 100-, and 1000-pg). The Clontech SMARTer v4 kit (Takara Bio USA, Inc., Mountain View, CA) was used for global preamplification of these serially diluted samples in triplicate and also for direct global preamplification in FACS-derived murine DG cell types prior to library generation in quadruplicates with the Nextera XT DNA Library Prep Kit (Illumina, Inc., San Diego, CA). Samples were sequenced to a depth of 15–20 million 2 × 150 bp clusters on the Illumina HiSeq 4000 platform (Illumina, Inc., San Diego, CA).

Publicly available data retrieval

Published data was retrieved in FASTQ format from the following accession numbers: ArrayExpress E-MTAB-2600 [32] (mouse embryonic stem cells from 2i and alternative 2i media) and GSE50856 [14] (SMART-seq samples).

Data preprocessing, alignment, and quantification

Individual FASTQ files were trimmed for adapter sequences and filtered for a minimum quality score of Q20 using AdapterRemoval v2.2.0 [33]. Preliminary alignment using HISAT2 v2.0.6 [34] was performed to a composite reference of rRNA, mtDNA, and PhiX bacteriophage sequences obtained from NCBI RefSeq [35]. Reads aligning to these references were excluded in subsequent analyses.

Walker *et al. J Transl Med*   (2020) 18:63

Page 4 of 15

Primary alignment was performed against the human genome reference GRCh38p7 or mouse genome reference GRCm38p4 using HISAT2. Gene expression values for genes described by the GENCODE [36, 37] Gene Transfer Format (GTF) release 25 (human) or release M14 (mouse) were quantified using the featureCounts tool of the Subread package v1.5.1 [38, 39] in unstranded mode.

Sequencing and alignment quality control
Quality control was performed using a modification of our custom workflow QuaCRS [40]. In brief, aligned read quality was verified using RNA-SeQC [41] and RSeQC [42]. Parameters evaluated included the exonic rate (the percentage of reads aligning to exons), the intronic rate (the percentage of reads aligning to introns), and the duplication rate (the percentage of reads that were identified as PCR duplicates).

Coverage profiling
Coverage depths across the aligned reference were calculated with a per-base resolution using the 'genomecov' utility of Bedtools v2.27.0 [43] in BedGraph format. It is imperative to utilize 'split output' mode to reduce the size of the output BedGraph files. These files are used as input into CLEAR.

Calculation of transcript μ_i
For each transcript (i) annotated in the NCBI RefSeq (GRCh38 or GRCm38) reference, as sourced from the UCSC Genome Browser [44], CLEAR calculates the transcript's μ_i parameter. This quantifies the distribution of the positional mean of the read distribution along that transcript between the 5′ ($\mu_i = -1$) and the 3′ ($\mu_i = +1$) ends (Eq. 1):

$$\mu_i = \frac{2}{L_{\text{transcript}}} \left(\frac{\sum_{k=0}^{L_{\text{transcript}}-1} (k \cdot d_k)}{\sum_{k=0}^{L_{\text{transcript}}-1} (d_k)} \right) - 1 \qquad (1)$$

where $L_{\text{transcript}}$ is the length of a given transcript, and d_k is the coverage of exonic locus k zero indexed and starting at the transcription start site. In the case that a gene contains multiple isoforms, the longest transcript from the UCSC genome browser is used for the μ_i calculation.

Determination of analysis-ready CLEAR transcripts
All transcripts quantified by featureCounts are sorted by overall length-normalized expression. Histograms of μ_i values from 250 transcripts each, are collected and fit using the 'optimize' module of the Python scipy package, to a double-beta distribution as described by Eq. 2:

$$H \left(\frac{(1+x)(1 - \frac{1}{2}(1+x))^{1+b}}{2B(2, 2+b)} + \frac{(1+x)^{1+a}(1 - \frac{1}{2}(1+x))}{2^{1+a}B(2+a, 2)} \right) \qquad (2)$$

where H is a normalization parameter fixed by the bin sizes, $B(q,p)$ is the beta integral of q and p, and x is the bin location. The fitting parameters are a and b, which are each bounded to be non-negative. A value of 0 for both parameters represents a symmetric distribution and positive values represent progressively bimodal distributions to the positive (a), or negative (b) direction. These windows are advanced in units of 10 transcripts, reducing computational intensity. Once a window is found where a or b have a value greater than 2, the software exports transcripts with a measured expression higher than that bin's value for analysis. For multi-sample comparisons, transcripts are overlapped and only transcripts found in all samples are included for downstream analysis. We note, however, that there is a command line argument available to reduce this requirement to only a subset of samples for investigators interested in a less stringent inclusion criteria.

CLEAR visualizations
A core component to the quality control of the CLEAR selection process is the visualization of μ_i to confirm that the characteristic bifurcation is observed. Violin plots are produced using code included with CLEAR which utilizes the matplotlib package in Python. Examples of this can be seen in Fig. 2d, Additional file 1: Figures S5A, C.

DEG comparisons
Differentially expressed genes were called using DESeq2 [18] run on counts tables generated with featureCounts as described above. In all summarization figures, a false discovery rate (FDR) q-value of < 0.05 was used as an inclusion criterion for DEGs.

Principal component analysis
Principal component analysis (PCA) was utilized to visualize differences between samples. All PCA plots were generated from counts tables that were size-normalized and r-log transformed after CLEAR selection using methods included with DESeq2. Each comparison was processed with the SciKitLearn [45] PCA implementation and plotted using custom scripts in Python.

Results
Developing CLEAR
The experimental design for developing CLEAR is presented in Fig. 1a. Our goal is to detect differences between two biological conditions using low input amounts and few replicates. To illustrate the feasibility

Walker *et al. J Transl Med*   (2020) 18:63

Page 5 of 15

Fig. 1 lcRNA-seq analyses between sample groups without application of CLEAR. **a** Workflow for total RNA extraction and QC analysis from FACS-derived CD5+ and CD5− CLL cells as input for lcRNA-seq library generation for the development of CLEAR; **b** the DEG counts as determined by DESeq2. Contrary to expectation, the 10-pg input has more DEGs than the 100-pg input replicates; **c** shared DEGs between the three input groups showing more DEGs that are unique than shared; **d** Unsupervised analysis of CD5+ and CD5− samples by total RNA input amount. PCA reveals sample replicates separated by biological groupings at the 1000- and the 100-pg input level but not at the 10-pg level. *CLL* Chronic lymphocytic leukemia, *PBMC* peripheral blood mononuclear cells, *FACS* fluorescence activated cell sorting, *BioA* agilent bioanalyzer RNA assay, *DEGs* differentially expressed genes, *PC* principal component

of this approach in a clinical study, we used total RNA extracted from FACS-derived primary CD5$^+$CD19$^+$ (CD5+) and CD5$^-$CD19$^+$ (CD5−) cells from a single chronic lymphocytic leukemia (CLL) patient to establish the workflow and custom analysis. The RNA was diluted to levels found in typical lcRNA-seq projects with three replicates per input mass (10-, 100-, and 1000-pg) and cell type. Although CLEAR is designed to identify robust transcripts in 'Direct Amplification' libraries prepared from a few cells, sample inputs used for its development were from extracted RNA and not from cell pools. The overall design rationale is as follows: (1) To track the efficacy of transcripts selected by CLEAR in downstream DESeq2 [18] analysis, the composition of input RNA had to be the same for all dilutions so that the only difference within each cell type was the input amount; (2) The 10- and 100-pg input levels approximate the RNA amount in ∼ 5 and ∼ 50 immune cells, respectively. Since systems noise from 1000-pg input level is considered acceptable [10], it serves as a 'gold standard' to evaluate the 10- and 100-pg transcriptomes.
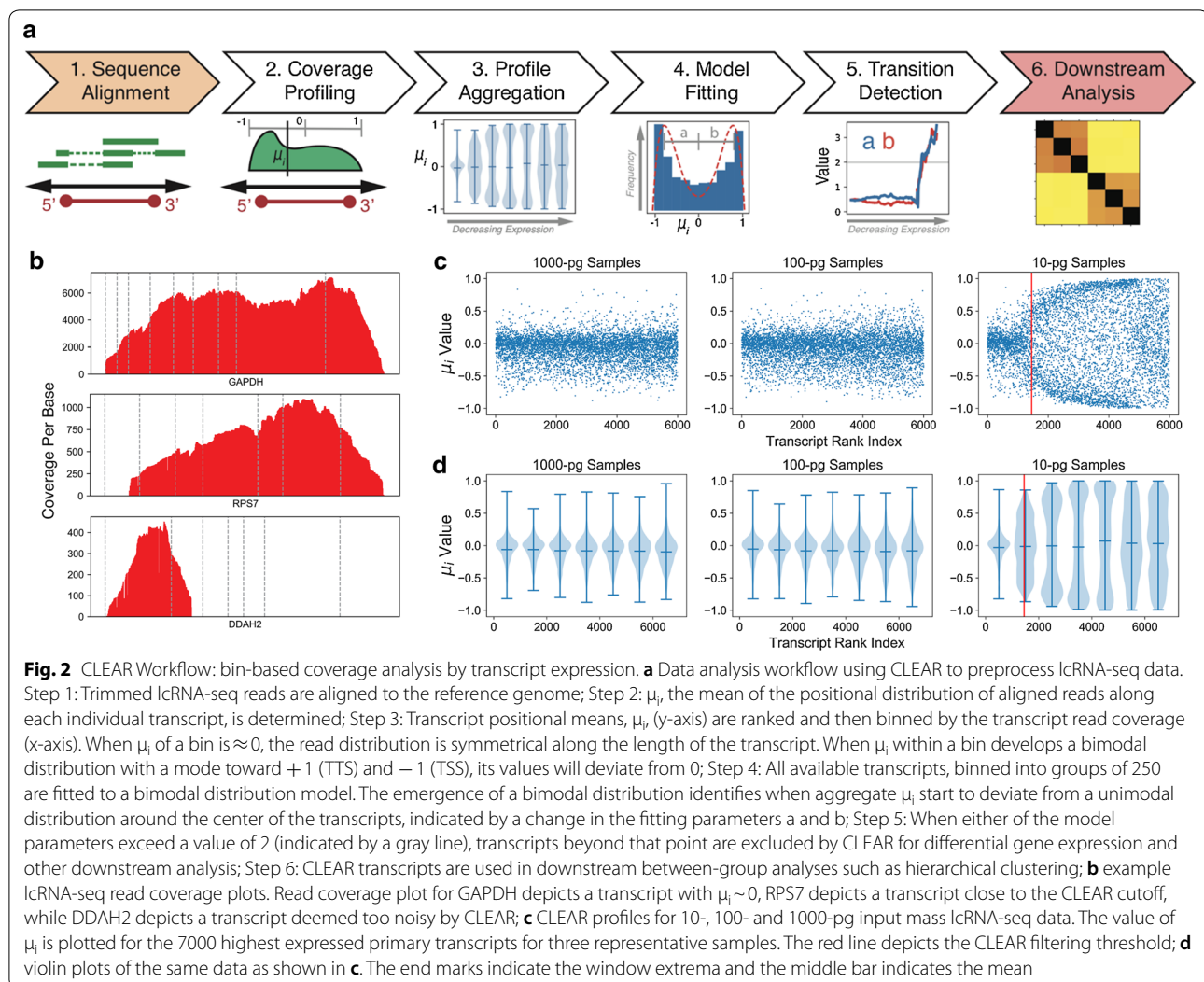
LcRNA-seq data quality from serial dilutions of the same RNA stock strongly depends on input mass and yields dissimilar DEGs when processed with a standard RNA-seq analysis pipeline

Average read depth was 15.5 million clusters, which is typical of bulk mRNA-seq experiments. Post-alignment

quality assessment was used to illustrate the effect of input RNA amounts on data quality. As expected, quality parameters correlate inversely with RNA inputs (Additional file 1: Figure S2), with the 10- and 100-pg samples yielding less transcriptomic information than the 1000-pg samples despite having similar sequencing depths.

To examine the effect of technical noise associated with lcRNA-seq experiments, we performed DEG analysis using a standard bulk cell RNA-seq workflow on the 10-, 100- and 1000-pg replicates without consideration of noise structure from signal dropouts and RNA degradation. As expected, the overall transcript variabilities, depicted in Additional file 1: Figure S3 as scatterplots between replicates of the two cell types, were highest in the 10-pg input RNA replicates. Within all samples, variabilities were highest in the low- to medium-expressed transcripts while the expression profiles converged at high-expressed transcripts.

Next, we performed DEG analysis using DESeq2 workflow without modifications [18]. DEG counts for the CD5+ vs. CD5− comparison were: 1000-pg: 2996, 100-pg: 744, and 10-pg: 898 (Fig. 1b). This outcome is unexpected as the statistical power to discern DEGs should increase with increasing RNA input mass. Of the DEGs identified between the two cell types, few were shared between any two input levels (Fig. 1c). This is also unexpected because the three input levels were diluted from the same RNA extraction stocks. In Fig. 1d, we present PCA plots

Walker *et al. J Transl Med*   (2020) 18:63

Page 6 of 15

Fig. 2 CLEAR Workflow: bin-based coverage analysis by transcript expression. **a** Data analysis workflow using CLEAR to preprocess lcRNA-seq data. Step 1: Trimmed lcRNA-seq reads are aligned to the reference genome; Step 2: μ_i, the mean of the positional distribution of aligned reads along each individual transcript, is determined; Step 3: Transcript positional means, μ_i, (y-axis) are ranked and then binned by the transcript read coverage (x-axis). When μ_i of a bin is ≈ 0, the read distribution is symmetrical along the length of the transcript. When μ_i within a bin develops a bimodal distribution with a mode toward $+1$ (TTS) and -1 (TSS), its values will deviate from 0; Step 4: All available transcripts, binned into groups of 250 are fitted to a bimodal distribution model. The emergence of a bimodal distribution identifies when aggregate μ_i start to deviate from a unimodal distribution around the center of the transcripts, indicated by a change in the fitting parameters a and b; Step 5: When either of the model parameters exceed a value of 2 (indicated by a gray line), transcripts beyond that point are excluded by CLEAR for differential gene expression and other downstream analysis; Step 6: CLEAR transcripts are used in downstream between-group analyses such as hierarchical clustering; **b** example lcRNA-seq read coverage plots. Read coverage plot for GAPDH depicts a transcript with $\mu_i \sim 0$, RPS7 depicts a transcript close to the CLEAR cutoff, while DDAH2 depicts a transcript deemed too noisy by CLEAR; **c** CLEAR profiles for 10-, 100- and 1000-pg input mass lcRNA-seq data. The value of μ_i is plotted for the 7000 highest expressed primary transcripts for three representative samples. The red line depicts the CLEAR filtering threshold; **d** violin plots of the same data as shown in **c**. The end marks indicate the window extrema and the middle bar indicates the mean

for cell type comparisons at the three input levels. At the 10-pg input amount, the gene expression profiles from the three CD5− replicates were vastly different from each other resulting in no separation between the cell types along either Principal Component 1 (PC1) or PC2. At the 100- and 1000-pg levels, the transcriptome profiles distinguished the two cell types, with PC1 accounting for 26.3% and 74.6% of the variance, respectively.

Taken together, the transcriptome profiles, the DEG dissimilarities and quantities, and the inconclusive PCA outcome reveal a high degree of systems noise at the 10-pg input level. Though systems noise was reduced at the 100- and 1000-pg input amounts, they still shared relatively few DEGs between them. As such, lcRNA-seq data clearly require custom analysis to control for systems noise.

Individual gene coverage profiles can guide selection of robust and reliably quantifiable transcripts in high noise conditions

It is known that systems noise associated with lcRNA-seq data are highest in lowly-expressed transcripts and manifested as large numbers of duplicated reads piling up in section(s) of these transcripts [14, 46]. As our aim is to provide a DEG analysis workflow for studies/samples with degraded RNA (e.g., multi-step cell-type enrichment from stored samples), low RNA content (e.g., < 100 immune cells), and few replicates, we elect to identify reliable transcripts in each replicate within each experimental group for between-group comparisons. The CLEAR workflow is presented in Fig. 2a and in "Methods".

In Fig. 2b, we illustrate CLEAR transcript evaluation using coverage profiles from three genes in a 10-pg CD5+ sample. The first, GAPDH, is a highly-expressed
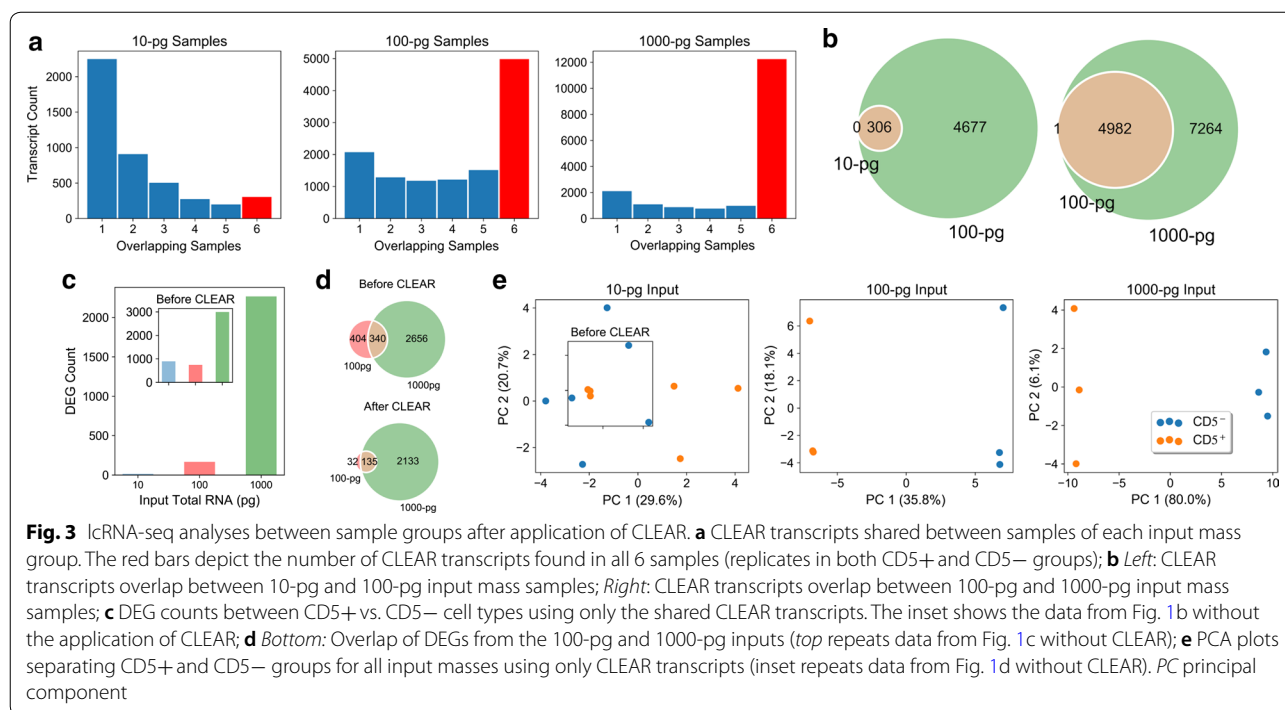
Walker *et al. J Transl Med*   (2020) 18:63

Page 7 of 15

housekeeping gene. This group of genes has uniform 5′ to 3′ read coverage, similar to coverage profiles in bulk RNA-seq. The second gene, RPS7, is in the last accepted CLEAR transcript bin. At this point, coverage profiles switch from an acceptable unimodality centering at the middle of the transcript to an unacceptable amount of bi-modality with region(s) of read pile-ups and region(s) with low/no coverage. Lastly, DDAH2 is below the CLEAR threshold. Here, the profile is dominated by regions with drastically different coverage. These profiles are prone to erroneous gene expression quantifications and false DEGs.

To illustrate this change in coverage profile associated with transcript integrity, we plot the coverage profile mean μ_i from CD5+ samples (one replicate per input level) in Fig. 2c, d as scatter- and violin plots, respectively. The 100- and 1000-pg samples have even read distributions ($\mu_i \approx 0$) for the 6000 highest expressed transcripts while the 10-pg transcripts experience unevenness distributions after the $2,000^{th}$ highest expressed transcript. The red line signifies the transition of transcripts passing CLEAR to transcripts too noisy for DEG analysis. Due to higher input amounts in the 100- and 1000-pg samples, their acceptable-to-unacceptable transitions occur at lower expression levels not shown in the plots (see discussion below).

CLEAR selection increases shared DEGs between CD5+ and CD5− samples at all RNA input levels and improves cell type separation at the 10-pg level

Earlier in this report (Fig. 1b–d; Additional file 1: Figure S3), we presented the impact of systems noise associated with lcRNA-seq libraries. In Fig. 3 and Additional file 1: Figures S4, S5, we present evidence for the transformative power of removing noisy transcripts from lcRNA-seq DEG analysis. First, we examine the impact of CLEAR on the number of robust transcripts per input level. The red histogram (Fig. 3a, *right* plot) shows 12,246 shared transcripts passing CLEAR in all 1000-pg samples (three replicates/cell type). The number drops to 4983 for the 100-pg samples and to 306 for the 10-pg samples. The high number of shared CLEAR transcripts among the 100- and 1000-pg samples signifies low system noise, while few CLEAR transcripts are shared among the 10-pg samples. This highlights that input amounts ≤ 5 cells are susceptible to systems noise. The same trends are observed when the shared CLEAR transcripts are investigated separately by cell type (Additional file 1: Figure S4).

The effects of CLEAR on replicate-to-replicate comparisons are presented in Additional file 1: Figure S5. The main takeaway is that CLEAR removes a large number of lowly-expressed transcripts from the 10-pg samples leaving highly-expressed transcripts for downstream analyses. The amount of retained transcripts increases with increasing RNA input. For the 1000-pg samples, we

Fig. 3 lcRNA-seq analyses between sample groups after application of CLEAR. **a** CLEAR transcripts shared between samples of each input mass group. The red bars depict the number of CLEAR transcripts found in all 6 samples (replicates in both CD5+ and CD5− groups); **b** *Left*: CLEAR transcripts overlap between 10-pg and 100-pg input mass samples; *Right*: CLEAR transcripts overlap between 100-pg and 1000-pg input mass samples; **c** DEG counts between CD5+ vs. CD5− cell types using only the shared CLEAR transcripts. The inset shows the data from Fig. 1b without the application of CLEAR; **d** *Bottom:* Overlap of DEGs from the 100-pg and 1000-pg inputs (*top* repeats data from Fig. 1c without CLEAR); **e** PCA plots separating CD5+ and CD5− groups for all input masses using only CLEAR transcripts (inset repeats data from Fig. 1d without CLEAR). *PC* principal component

Walker *et al. J Transl Med*   (2020) 18:63

Page 8 of 15

observe transcripts from the entire transcription range passing CLEAR. This supports ours and others' postulation [10] that the molecular complexity in 1 ng RNA approaches that found in bulk RNA-seq. We note subtle differences between each comparison pair, illustrating that CLEAR assesses samples independently to derive non-static cutoff values.

The Venn diagrams in Fig. 3b show the overlaps of CLEAR transcripts (depicted by the red bars in Fig. 3a) common to all 6 samples (3 replicates/cell type) per input level. With CLEAR successfully excluding noisy transcripts from these lcRNA-seq data, nearly all of the shared CLEAR transcripts at each input level are contained in the shared CLEAR transcripts at the respective higher input level.

Similarly, CLEAR pre-selection positively impacts the DEG analysis between cell types (Fig. 3c, d). In contrast to using all transcripts (Fig. 1b; also Fig. 3c *inset*), DEGs are now lowest (n = 3) in the 10-pg comparison, intermediate (n = 189) in the 100-pg comparison, and largest (n = 2826) in the 1000-pg comparison. This trend is expected as power to detect DEGs increases with increasing RNA input. Also, CLEAR pre-selection alters the DEG overlap shown in Fig. 1c. When only CLEAR transcripts are used (*bottom* Fig. 3d), the 3 DEGs at the 10-pg inputs are: CD69 (found in all comparisons); SRSF11 (also found in the 1000-pg comparison); and HINT1 (only found in the 10-pg comparison). Of the 189 DEGs at the 100-pg input level, 135 are shared with the 1000-pg input. These overlaps far exceed DEGs derived from all transcripts (Fig. 1c and *top* Fig. 3d).

Finally, applying CLEAR also improves PCA segregation of cell types (Fig. 3e). In contrast to results from data without CLEAR selection (Fig. 1d), the CD5+ replicates are now separated from the CD5− replicates along PC1 at all input levels (Fig. 3e) and not just the 100- and 1000-pg levels (Fig. 1d).

To conclude, we remind readers that the dilution samples were prepared from the same RNA stock, the library generation was performed using the same reagents, by the same scientist, at the same time, and sequenced on the same flow cell. The dramatic differences in data quality observed in Fig. 2c, d are mainly associated with the input RNA amount. While the traditional analysis shown in Fig. 1 is severely hampered by the amount of systems noise in lcRNA-seq data, implementation of CLEAR results in overlapping and convergent DEGs at all input levels.

The number of CLEAR transcripts correlates with authors' quality measures in public scRNA-seq data

To evaluate CLEAR's utility in identifying robust transcripts in published ultralow-input full-length RNA-seq data, we selected an scRNA-seq dataset based on the Fluidigm/SMART-seq approach. Ilicic et al. [25] provided detailed quality characteristics on captured single cells. We selected the mouse embryonic stem cells (576 cells grown in 2i or alternative 2i media) to compare CLEAR's assessments of the captured cells vs. the authors' quality calls of 'Good' and 'Empty/No capture'. In Additional file 1: Figure S6A, we present violin plots illustrating μ_i from scRNA-seq data from a 'Good' cell. We note that the coverage bifurcations observed in our lcRNA-seq data are also observed in the scRNA-seq data. In Additional file 1: Figure S6B, we provide CLEAR transcript count distributions in the two quality-groups. For the 521 'Good' cells, they have similar CLEAR transcript counts with mean of 4098. The 'Empty' cells, however, have widespread CLEAR transcript counts with mean at 651 and with 10 of 15 cells having 0. As depicted, the mean value is skewed by two outliers, potentially due to ambient RNA from burst cells prior to imaging.

CLEAR transcript numbers are similar between studies with similar input RNA mass

Next, we applied CLEAR to SMART-Seq data from Bhargava et al. [14] prepared from serial dilution (25-, 50-, 100- and 1000-pg) of polyA-selected mRNA from control and Activin A-treated mouse embryonic stem cells. Again, we note that the read coverage bifurcations observed in our lcRNA-seq data also occurred here (Additional file 1: Figure S6C). Additional file 1: Figure S6D presents CLEAR transcript counts for the four samples (control and treated groups, two replicates each) at all input levels. As expected, CLEAR transcript counts increase with increasing mRNA inputs. When comparing CLEAR transcript counts in the Bhargava et al. [14] data to our CD5+/CD5− data, it is important to note that different input material was used (polyA-selected mRNA in the former and total RNA in the latter) and associated differences in sample preparation chemistry. To that end, we apply a conservative adjustment factor of 10 [47] between total RNA and mRNA. As such, our 1000-pg total RNA 'gold standard level' is comparable to Bhargava et al.'s [14] 100-pg mRNA input. Indeed, the Bhargava et al. [14] 100-pg mRNA libraries produced from Activin A stimulated cells had similar shared CLEAR transcript counts (mean, 14,940) to our 1000-pg CD5+/CD5− replicates (mean, 14,717). It is worth noting, however, that the unstimulated control samples from the Bhargava study produced noticeably fewer CLEAR transcripts (mean, 6735), demonstrating the ability of CLEAR to distinguish active- and quiescent state between samples. At the intermediate input level of 100-pg total RNA, the closest corresponding level for the Bhargava et al. [14] study would be 10-pg mRNA. Since their lowest

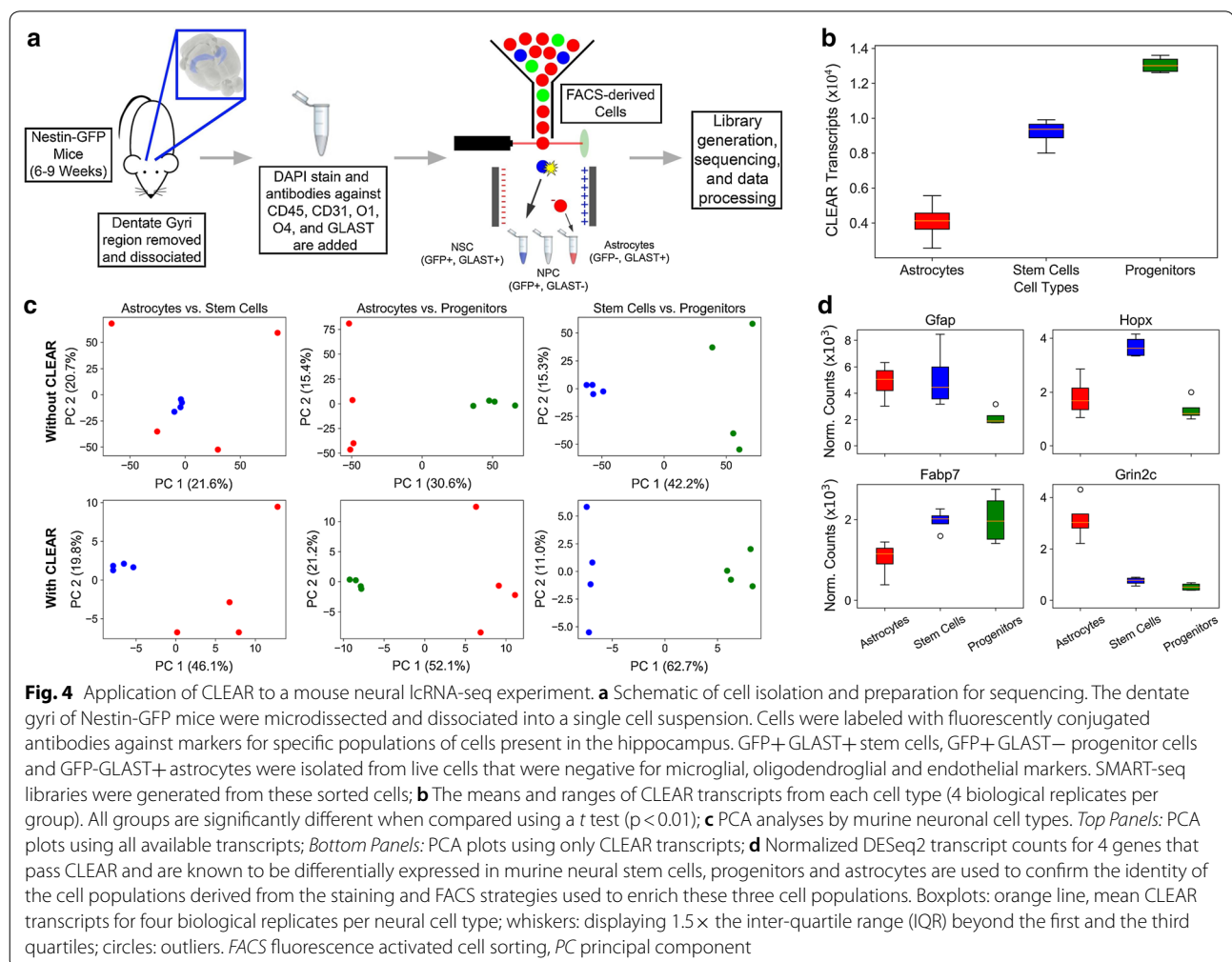Walker *et al. J Transl Med*   (2020) 18:63

Page 9 of 15

input level was 25-pg, it makes sense that their average CLEAR transcript counts of 6408 are similar to our average transcript counts of 8420 at 100-pg input. Due to the much higher RNA input amount, the data in the Bhargava et al. [14] study was not challenging enough to highlight CLEAR's ability to improve DEG and PCA analyses between the two biological groups, in agreement with the authors' original assessment that standard differential expression methods could be applied.

CLEAR improves group separation of murine neural cell types and recapitulates genes known to be differentially expressed between them

To demonstrate the applicability of CLEAR in a biological system, we applied it to lcRNA-seq data derived from three cell types in murine hippocampal dentate gyrus (DG), a highly-studied neurogenic niche associated with memory and cognitive function [48]. This region is of particular interest because resident stem and glial cells show robust morphological and proteomic responses

to a variety of injuries such as trauma and seizures [49, 50]. However, the stem cells and local glial populations are small cell populations embedded in a heterogeneous niche making them perfect candidates for FACS-derived lcRNA-seq analysis.

The study schematics are shown in Fig. 4a and the CLEAR transcript counts in Fig. 4b. The data reveal that neural progenitor cells had the highest number of CLEAR transcripts, followed by neural stem cells and then astrocytes. While this trend may reflect differences in cell size and RNA content, it may also reveal the tolerance of each cell type toward cell dissociation and enrichment conditions used. For example, astrocytes possess a complex branching of fine processes extending from their cell body that could be damaged during the mechanical dissociation and flow sorting. Furthermore, they are intrinsically reactive to changes in the external environment [51] such as the chemical dissociation step and the FACS media required by the Clontech SMARTer kit

Fig. 4 Application of CLEAR to a mouse neural lcRNA-seq experiment. **a** Schematic of cell isolation and preparation for sequencing. The dentate gyri of Nestin-GFP mice were microdissected and dissociated into a single cell suspension. Cells were labeled with fluorescently conjugated antibodies against markers for specific populations of cells present in the hippocampus. GFP+ GLAST+ stem cells, GFP+ GLAST− progenitor cells and GFP-GLAST+ astrocytes were isolated from live cells that were negative for microglial, oligodendroglial and endothelial markers. SMART-seq libraries were generated from these sorted cells; **b** The means and ranges of CLEAR transcripts from each cell type (4 biological replicates per group). All groups are significantly different when compared using a *t* test (p < 0.01); **c** PCA analyses by murine neuronal cell types. *Top Panels:* PCA plots using all available transcripts; *Bottom Panels:* PCA plots using only CLEAR transcripts; **d** Normalized DESeq2 transcript counts for 4 genes that pass CLEAR and are known to be differentially expressed in murine neural stem cells, progenitors and astrocytes are used to confirm the identity of the cell populations derived from the staining and FACS strategies used to enrich these three cell populations. Boxplots: orange line, mean CLEAR transcripts for four biological replicates per neural cell type; whiskers: displaying 1.5× the inter-quartile range (IQR) beyond the first and the third quartiles; circles: outliers. *FACS* fluorescence activated cell sorting, *PC* principal component

(phosphate-buffered saline devoid of fetal bovine serum, Ca^{2+}, and Mg^{2+}).

Despite suboptimal cDNA quality from the astrocytes, the application of CLEAR greatly enhanced cell-type separations when compared to using all available transcripts (Fig. 4c). When all transcripts were used, PCA plots show clear separation between stem cells and progenitors (third panel on *top*), but the lower transcript quality in astrocytes renders the discrimination between astrocytes and stem cells unclear (first panel on *top*). The astrocyte transcript quality does not impact its separation from the progenitors due to large differences in transcript profiles between these two cell types (second panel on *top*). Upon application of CLEAR, astrocytes and stem cells become well separated (first panel on *bottom*) while retaining separation between astrocytes and progenitors (second panel on *bottom*).

Genes known to be differentially expressed in the three cell types (Table 1) are used to confirm the efficacy of the enrichment strategies and of the CLEAR analysis. Of the transcripts listed in Table 1, only four shown in the top section and in Fig. 4d passed CLEAR in all three cell types; several others pass in only two cell types, allowing differential expression only between those two. Genes that failed CLEAR are presented in Additional

file 1: Figure S7. Glial fibrillary acidic protein (Gfap) and SRY-Box 9 (Sox9) are known to be higher in astrocytes and stem cells than in progenitors, which we also found (q < 0.01 and 0.0001, respectively). Glutamate ionotropic receptor NMDA type subunit 2C (Grin2c) and HOP homeobox (Hopx) are known to be different between stem cells and astrocytes, with Hopx higher in stem cells and Grin2c higher in astrocytes. Our findings confirmed both trends (q < 0.01 and q < 0.001, respectively). In addition, our data show that the expressions of Hopx and Grin2c are significantly higher in stem cells and astrocytes than in progenitors (q < 0.0001 for both). Finally, the expression of fatty acid binding protein 7 (Fabp7) is known to be higher in progenitors and stem cells relative to astrocytes as confirmed in our experiment (q < 0.01 and q < 0.05, respectively).

For genes that do not pass CLEAR, the picture is more nuanced. We examine normalized count plots for a few of the known discriminating transcripts (Additional file 1: Figure S7). Inhibitor of DNA binding 4 (Id4) and SRY-Box 2 (Sox2) are known to have similar transcript levels in astrocytes and stem cells but higher than those in progenitors, which is what we found. Minichromosome maintenance complex component 2 (Mcm2), doublecortin (Dcx), eomesodermin (Eomes), proliferating cell nuclear antigen (Pcna), neuronal differentiation 1 (Neurod1), and neurogenin 2 (Neurog2) have higher expressions in the progenitors [52–57], which we also find. However, nestin (Nes) is expected to be expressed in both stem cells and progenitors, but is only observed in the latter. Without CLEAR, Nes would be called as significantly higher in progenitors (q ≤ 0.01). With CLEAR, Nes fails its criteria thereby avoiding a false finding. Taken together, our dataset recapitulates and reinforces relative gene expression patterns reported for neural stem cells, progenitors, and astrocytes and shows that staining, sorting, and lcRNA-seq with CLEAR can be used to distinguish these cell types.

Discussion

Total RNA input amount for lcRNA-seq (e.g., 10–1000-pg from 1 to 100 cells) is closer to amounts found in scRNA-seq studies (1–10-pg) than to bulk RNA-seq studies (100–200-ng). Due to the minute RNA input amount in single-cell/limiting-cell transcriptomic studies, the library generation process, with the exception of single-molecule sequencing, is preceded by global cDNA preamplification. Researchers have observed artefacts such as a large proportion of the sequencing reads being dominated by a small number of transcripts [14], excessive transcripts with zero read counts [58], high variabilities between and within replicates of sample groups [30], distortion towards shorter transcripts [59],

Table 1 Application of CLEAR and DESeq2 to murine DG cell type comparisons

Gene	Astrocyte vs. progenitor	Progenitor vs. stem cell	Astrocyte vs. stem cell
Gfap	N.S.	**	N.S.
Hopx	N.S.	****	**
Fabp7	**	N.S.	*
Grin2c	****	N.S.	***
Sox9	F.C.	****	F.C.
Neurod1	F.C.	F.C.	F.C.
Dcx	F.C.	F.C.	F.C.
Id4	F.C.	***	F.C.
Pcna	F.C.	N.S.	F.C.
Mcm2	F.C.	F.C.	F.C.
Ascl1	F.C.	F.C.	F.C.
Eomes	F.C.	F.C.	F.C.
Nes	F.C.	F.C.	F.C.
Neurog2	F.C.	F.C.	F.C.

Genes known to be differentially expressed in murine astrocytes, stem cells and progenitors. lcRNA-seq data were preprocessed using CLEAR prior to between-group comparisons using DESeq2. Differential gene expression analysis evaluates the effectiveness of staining and FACS strategies in enriching these three cell types. The bisecting line indicates the border between transcripts which pass CLEAR in all samples and those that do not pass CLEAR in all samples. Each comparison was processed using DESeq2 and the significance of the comparison is given. *q:* FDR-corrected p-values; N.S. (Not Significant), F.C. (Failed CLEAR), * (q ≤ 0.05), ** (q ≤ 0.01), *** (q ≤ 0.001), **** (q ≤ 0.0001)

Walker *et al. J Transl Med*   (2020) 18:63

Page 11 of 15

and higher variances at lower biological abundances [10, 11]. Together, these artefacts result in noisy sc/lcRNA-seq data that challenge assumptions fundamental to bulk RNA-seq analysis approaches and render them unsuitable for direct application to these data [16]. Existing strategies to overcome challenges associated with noisy data to uncover biological differences in single-cell studies include: (1) identifying low-quality, high-noise samples and removing them from downstream analyses [25]; (2) applying transcript normalization [16, 21, 27, 60]; (3) incorporating ERCC spike-in control [30, 61]; (4) utilizing median absolute deviation (more resistant to outliers) to characterize statistical dispersion [12]; (5) integrating UMIs to track transcripts by molecular counts [62]. Yet, due to the often low number of replicates and the goal of identifying DEGs between sample groups, the strategies for scRNA-seq data just described are not appropriate or adequate for lcRNA-seq data.

Increasingly, RNA-seq is used in clinical trials to study patient's transcriptional response to novel drugs or drug combinations. Patient biospecimens are extremely precious, especially for well annotated samples. In this scenario and in cases where cell types of interest are rare, the combination of lcRNA-seq (data generation) and CLEAR (identification of robust transcripts from high noise data for between-group comparisons) enables global profiling without exhausting important samples while preserving RNA quality by bypassing total RNA extraction. Similar to single-cell analysis, lcRNA-seq has high signal dropouts and skewed read coverage profiles. CLEAR is designed to identify transcripts with acceptable systems noise for downstream analyses by evaluating one sample at a time (not by assigning a uniform cutoff across all samples). CLEAR then requires transcripts to be reliably quantifiable in all replicates within and across comparison groups. This criterion is made possible due to the RNA quality preservation effects of direct cDNA preamplification reagents and library generation using nanoliter-microfluidics devices as they increase even read coverage in lcRNA-seq transcripts. To our knowledge, preprocessing lcRNA-seq data using CLEAR followed by DEG analysis is unique and unlike approaches used by others in similar studies. For example, Shanker et al. [13] performed Pearson correlation comparisons between serial dilutions of input RNA vs. their 1 μg input control using $\log_2$ of median RPKM for all genes having ≥ 2 reads coverage. Bhargava et al. [14] performed DESeq without prior data preprocessing to compare the performance between three library preparation methods as well as within serial dilutions of mRNA derived from two mouse embryonic stem cell culture conditions. Liu et al. [63] utilized relative expression orderings to harmonize clinical transcriptome signatures between low-input

and bulk RNA-seq libraries. In contrast, CLEAR comprehensively screens all transcripts in all replicates within an lcRNA-seq study for genes with even coverage for downstream analyses. The combination of characterizing each transcript by its read distribution mean μ_i followed by cutoff selection defined by expression across the full transcript distribution is deliberate. For any given transcript, even one with low systems noise, there are many reasons why its mean μ_i could deviate from zero: (1) transposase-based library generation methods [64] less effectively 'tagment' the 5′-end of transcripts (3′-ends are less affected likely due to the presence of poly(A) tails), affecting 5′-end read coverage; (2) transcript isoforms skew the read distribution along the length of the transcript, especially when prominent isoforms are not correctly annotated [17]; (3) presence of truncated, polyadenylated RNA fragments/intermediates part way through the RNA decay pathway [65]; (4) presence of RNA fragments with alternative polyadenylation sites [17, 66, 67]. Thus, a non-zero μ_i alone does not signify a noisy transcript. Conversely, read coverage profiles of some of the noisy transcripts can become random and μ_i can approach zero by chance. Simply keeping transcripts with a mean μ_i close to zero would result in some valid transcripts being excluded and some noisy transcripts being included. By characterizing the entire distribution of μ_i over a defined range of expression levels, outliers due to the above-mentioned effects do not impact the classification of that particular expression bin. E.g., a well quantified transcript with an incorrectly annotated isoform would have a non-zero μ_i but would be found in a bin with a majority of genes with μ_i close to zero and thus classified as reliably quantifiable. In this manner, CLEAR adapts to the quality of each sample to assess systems noise, an approach more precise than using an arbitrary cutoff. It is, however, important to note that systematic biases along transcripts that affect *all* genes, such as the known 3′ bias of the original SMART-seq protocol [6] *do* negatively affect CLEAR; we thus recommend its use only on the SMART-seq2 and newer protocols.

In this report, we have selected two sets of public data [14, 25] and one set of in-house data to illustrate the utility of CLEAR. One of the goals for using public datasets was to ascertain that the mean read/fragment distribution phenomenon observed in the CLEAR development data is also present in sc- and lcRNA-seq libraries derived from Fluidigm C1 IFCs and manual SMART-seq, respectively. As expected, the violin plots from the public data (Additional file 1: Figure S6A, C) illustrate the distribution of μ_i transitions from an acceptable distribution centered around 0 to an unacceptably broad/bimodal distribution. Another goal was to evaluate the ability of CLEAR to identify noisy

Walker *et al. J Transl Med*   (2020) 18:63

Page 12 of 15

transcripts in a broad range of datasets. The application of CLEAR to Ilicic et al. scRNA-seq data [25] revealed association between cell integrity and the resultant data quality. Additional file 1: Figure S6B displays violin plots of the dispersion of CLEAR transcripts which recapitulated the authors' microscopy assessments of intact 'single cells' and 'no cell'. When applied to Bhargava et al. data [14, 25], CLEAR identified increasing amounts of 'passed' transcripts with increasing mRNA input mass (Additional file 1: Figure S6D) which we also observe in the CD5+/CD5− data. Together, these assessments confirm that CLEAR can independently reveal sample/transcript quality and identify noisy data associated with empty wells (Ilicic et al. data [14, 25]) and increases in data quality with increasing levels of RNA input (Bhargava et al. data [14, 25]). As yet another example of CLEAR's performance, we incorporated a proof-of-principle experiment with cells isolated from the neurogenic niche in murine hippocampus. Traditionally, this study has been hampered by the limited numbers of stem cells and progenitors present in murine DG. By coupling preamplification-based lcRNA-seq with CLEAR, DEGs between these cell types can be achieved with only one mouse brain per biological replicate.

When comparing different biological groups it is possible that a gene is highly expressed in one but not in the other. An example of this is depicted in the Eomes gene in Additional file 1: Figure S7. It passes CLEAR in the progenitors but fails in the other cell types. While true DEG analysis is not possible for such genes, one can still use the CLEAR cutoff expression level (the count of the lowest transcript passing CLEAR for a particular sample) to report a bound on the fold change of such a gene consistent with the CLEAR analysis. For example, DESeq2 reports a $\log_2$ fold change of 8.97 when comparing the expression of Eomes of progenitors to astrocytes. However, the 'bound' procedure suggests a more modest difference of 1.48, when substituting the threshold for the featureCounts-reported expression value. The latter analysis is likely more reliable than the higher fold change reported by DESeq2 alone due to the presence of signal dropouts when all transcripts are used. For large enough numbers of replicates, it also may be reasonable to identify a gene as a DEG (albeit without quantifying its fold change) if it passes CLEAR in all samples of one group and fails CLEAR in all samples of the other group as such behavior is exceedingly unlikely to occur by chance in the case of many replicates.

Although CLEAR can be applied to scRNA-seq data just as effectively as lcRNA-seq data (Additional file 1: Figure S6A, B, results from analyzing Ilicic et al. [25] scRNA-seq data using CLEAR), the low complexity and highly variable nature of scRNA-seq data would typically result in insufficient CLEAR transcripts for meaningful DEG analysis, especially when requiring transcripts to pass the CLEAR criterion in all (usually of a very large number) cells. The incorporation of imputation would relax the CLEAR criterion thereby rendering it useful in scRNA-seq data.

Conclusion

In conclusion, utilizing a direct cDNA amplification approach, the lcRNA-seq strategy allows researchers to perform global transcriptome profiling and identify DEGs using a few cells. We develop the CLEAR algorithm around two prevalent phenomena in ultralow input RNA-seq data: (1) systems noise due to global preamplification of low RNA amount is more prevalent in medium- to lowly expressed transcripts; (2) the read coverage profile of noisy transcripts contain discernible coverage gaps instead of even coverage along the transcript length. By gauging the average read coverage profile mean μ_i in transcripts binned by gene expression levels, CLEAR is able to systematically identify transcripts with sufficient integrity in all replicates across comparison groups for downstream analyses. Lastly, we note that by displaying μ_i as violin plots, CLEAR is useful as a visual QC tool to eliminate or set aside data from scRNA-seq samples that fail at the cell enrichment stage or during library generation. In all, we highlighted the functionalities CLEAR brings to sc/lcRNA-seq data preprocessing and data quality visualization and hope to spark additional computational and statistical development in this area.

Supplementary information

Additional file 1: Figure S1. FACS parameter diagrams for the enrichment of CD5+ and CD5− cells from a CLL patient PBMC sample. FACS flow diagrams depicting the steps involved in the enrichment of CD5+ cells and CD5− cells from live CD45+, CD3−, CD33−, CD19+ cells, followed by separation into either CD5+ or CD5− collection tubes (Fig. 1). FACS: Fluorescence Activated Cell Sorting. **Figure S2.** Survey of quality control (QC) metrics of RNA-Seq data. All sequencing was subject to quality control as described in "Methods". Key metrics are summarized here. Notably, in many metrics such as Intergenic Rate, Alignment Rate, and Duplication Rate, the 10-pg groups indicate lower quality libraries than 100-pg and 1000-pg. "Top30" corresponds to the proportion of reads that belong to the 30 highest genes by expression. Boxplots: orange line, mean metric value; whiskers: displaying 1.5× the inter-quartile range (IQR) beyond the first and the third quartiles; circles: outliers. **Figure S3.** Between-sample correlations of detected RNA-Seq read counts. Scatter plots are drawn comparing each sample to each other sample for each input mass. 10-pg samples show much more scattered counts, whereas 100-pg and 1000-pg samples show progressively higher correlation. **Figure S4.** Comparison of overlapping transcripts. The analysis from Fig. 3a was repeated, although CD5− and CD5+ samples were considered separately. Notably, the trend between CD5+ and CD5− mirrors that

of the pooled data in Fig. 3a. **Figure S5.** CLEAR Filtering results in fewer noisy transcripts at the 10-pg sample level. Analysis from Figure S3 was repeated using CLEAR-filtered gene counts. Notably, 10-pg samples are observed to be sparser, while the remaining data points are of much higher correlation. **Figure S6.** Application of CLEAR to public datasets. **A**, **B** data from Ilicic et al. [25] was processed using the CLEAR pipeline; **C**, **D** data from Bhargava et al. [14] was processed using the CLEAR pipeline; **A)** An example CLEAR trace from released data shows a representative separation; **B)** CLEAR transcript identity allows the separation of cells the authors classified as "Empty" from those classified as "Good." **C)** An additional example trace; **D)** CLEAR transcript counts are indicative of the input mRNA mass used to generate a sequencing library. **Figure S7.** Neuronal cell type markers which did not pass the CLEAR criterion. Similar to Fig. 4d, for each remaining gene, expression was plotted using the raw counts. Individual cell types which passed CLEAR filtering are indicated with an asterisk (*) below the respective box plot. Boxplots: orange line, mean CLEAR transcripts for four biological replicates per neural cell type; whiskers: displaying 1.5X the interquartile range (IQR) beyond the first and the third quartiles; circles: outliers.

Abbreviations
CD19: Cluster of differentiation 19; CD5: Cluster of differentiation 5; CLEAR: Coverage-based limiting-cell experiment analysis for RNA-seq; CLL: Chronic lymphocytic leukemia; DEG: Differentially expressed gene; DG: Dentate Gyrus; Dcx: Doublecortin; ERCC: External RNA Controls Consortium; Eomes: Eomesodermin; FACS: Fluorescence activated cell sorting; FDR: False discovery rate; Fabp7: Fatty acid binding protein 7; GTF: Gene transfer format; Gfap: Glial fibrillary acidic protein; Grin2c: Glutamate ionotropic receptor NMDA type subunit 2C; Hopx: HOP homeobox; Id4: Inhibitor of DNA binding 4; Mcm2: Minichromosome maintenance complex component 2; Nes: Nestin; Neurod1: Neuronal differentiation 1; Neurog2: Neurogenin 2; PC1: Principal component 1; PC2: Principal component 2; PCA: Principal component analysis; PDD: Papain/dispase/DNase; Pcna: Proliferating clear antigen; RNA: Ribonucleic acid; SMART-seq: Switching Mechanism At the 5′-end of the RNA Transcript sequencing; Sox2: SRY-Box 2; Sox9: SRY-Box 9; lcRNA-seq: Limiting-cell RNA-seq; mRNA: Messenger RNA; scRNA-seq: Single-cell RNA-seq.

Acknowledgements
None.

Software availability and requirements
Project Name: CLEAR.
    Project Home Page: https://github.com/rbundschuh/CLEAR.
    Operating system(s): Linux.
    Programming Language: Python.
    Other Requirements: Python (2.7.11 or above), Matplotlib (2.0 or above).
    License: GPL3.
    Any restrictions to use by non-academics: None.

Authors' contributions
Conceived and designed the experiments: LAW, EDK, JCB, NM, RB, PY. Performed the experiments: MGS, C-LC, EH, JKD, XC. Analyzed the data: LAW. Wrote the paper: LAW, EDK, JKD, RB, PY. All authors read and approved the final manuscript.

Availability of data and materials
All original sequencing files have been deposited to Gene Expression Omnibus (GEO) under accession numbers GSE115032 (human CD5+ and CD5− data) and GSE115033 (mouse neural data).

Ethics approval and consent to participate
All human samples were obtained following written informed consent in accordance with the Declaration of Helsinki and under a protocol reviewed and approved by the Institutional Review Board of the Ohio State University.

Consent for publication
Not applicable.

Competing interests
The authors declare that they do not have any competing interests.

Author details
[1] Department of Physics, College of Arts and Sciences, The Ohio State University, Columbus, OH, USA. [2] The Ohio State University Comprehensive Cancer Center, The Ohio State University, Columbus, OH, USA. [3] Division of Hematology, Department of Internal Medicine, College of Medicine, The Ohio State University, Columbus, OH, USA. [4] Department of Psychology, College of Arts and Sciences, The Ohio State University, Columbus, OH, USA. [5] Chronic Brain Injury Program, The Ohio State University, Columbus, OH, USA. [6] Department of Chemistry & Biochemistry, College of Arts and Sciences, The Ohio State University, Columbus, OH, USA. [7] Center for RNA Biology, The Ohio State University, Columbus, OH, USA.

References
1. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009;6:377–82.
2. Picelli S. Single-cell RNA-sequencing: the future of genome biology is now. RNA Biol. 2017;14:637–50.
3. Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, et al. Comparative analysis of single-cell RNA sequencing methods. Mol Cell. 2017;65(631–643):e4.
4. Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med. 2017;9:75.
5. Gupta I, Collier PG, Haase B, Mahfouz A, Joglekar A, Floyd T, et al. Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. Nat Biotechnol. 2018;36:1197–202.
6. Ramsköld D, Luo S, Wang Y-C, Li R, Deng Q, Faridani OR, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat Biotechnol. 2012;30:777–82.
7. Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods. 2013;10:1096–8.
8. Picelli S, Faridani OR, Björklund AK, Winberg G, Sagasser S, Sandberg R. Full-length RNA-seq from single cells using Smart-seq2. Nat Protoc. 2014;9:171–81.
9. Sasagawa Y, Nikaido I, Hayashi T, Danno H, Uno KD, Imai T, et al. Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. Genome Biol. 2013;14:R31.
10. Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, Myers RM, et al. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. Genome Res. 2014;24:496–510.
11. Streets AM, Zhang X, Cao C, Pang Y, Wu X, Xiong L, et al. Microfluidic single-cell whole-transcriptome sequencing. Proc Natl Acad Sci USA. 2014;111:7048–53.
12. Wu AR, Neff NF, Kalisky T, Dalerba P, Treutlein B, Rothenberg ME, et al. Quantitative assessment of single-cell RNA-sequencing methods. Nat Methods. 2014;11:41–6.
13. Shanker S, Paulson A, Edenberg HJ, Peak A, Perera A, Alekseyev YO, et al. Evaluation of commercially available RNA amplification kits for RNA sequencing using very low input amounts of total RNA. J Biomol Tech. 2015;26:4–18.

Walker *et al. J Transl Med*   (2020) 18:63

Page 14 of 15

14. Bhargava V, Head SR, Ordoukhanian P, Mercola M, Subramaniam S. Technical variations in low-input RNA-seq methodologies. Sci Rep. 2014;4:3678.

15. Adiconis X, Borges-Rivera D, Satija R, DeLuca DS, Busby MA, Berlin AM, et al. Comparative analysis of RNA sequencing methods for degraded or low-input samples. Nat Methods. 2013;10:623–9.

16. Vallejos CA, Risso D, Scialdone A, Dudoit S, Marioni JC. Normalizing single-cell RNA sequencing data: challenges and opportunities. Nat Methods. 2017;14:565–71.

17. Grün D, van Oudenaarden A. Design and analysis of single-cell sequencing experiments. Cell. 2015;163:799–810.

18. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.

19. Hao J, Cao W, Huang J, Zou X, Han Z-G. Optimal gene filtering for single-cell data (OGFSC)—a gene filtering algorithm for single-cell RNA-seq data. Bioinformatics. 2019;35:2602–9.

20. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018;36:411–20.

21. Vallejos CA, Marioni JC, Richardson S. BASiCS: Bayesian analysis of single-cell sequencing data. PLoS Comput Biol. 2015;11:e1004333.

22. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014;32:381–6.

23. McCarthy DJ, Campbell KR, Lun ATL, Wills QF. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinformatics. 2017;33:1179–86.

24. Angerer P, Haghverdi L, Büttner M, Theis FJ, Marr C, Buettner F. destiny: diffusion maps for large-scale single-cell data in R. Bioinformatics. 2016;32:1241–3.

25. Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, McCarthy DJ, Marioni JC, et al. Classification of low quality cells from single-cell RNA-seq data. Genome Biol. 2016;17:29.

26. Lin P, Troup M, Ho JWK. CIDR: ultrafast and accurate clustering through imputation for single-cell RNA-seq data. Genome Biol. 2017;18:59.

27. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. 2015;16:278.

28. Guo M, Wang H, Potter SS, Whitsett JA, Xu Y. SINCERA: a pipeline for single-cell RNA-seq profiling analysis. PLoS Comput Biol. 2015;11:e1004575.

29. van Dijk D, Sharma R, Nainys J, Yim K, Kathail P, Carr AJ, et al. Recovering gene interactions from single-cell data using data diffusion. Cell. 2018;174(716–729):e27.

30. Brennecke P, Anders S, Kim JK, Kołodziejczyk AA, Zhang X, Proserpio V, et al. Accounting for technical noise in single-cell RNA-seq experiments. Nat Methods. 2013;10:1093–5.

31. Mignone JL, Kukekov V, Chiang A-S, Steindler D, Enikolopov G. Neural stem and progenitor cells in nestin-GFP transgenic mice. J Comp Neurol. 2004;469:311–24.

32. Kolodziejczyk AA, Kim JK, Tsang JCH, Ilicic T, Henriksson J, Natarajan KN, et al. Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. Cell Stem Cell. 2015;17:471–85.

33. Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res Notes. 2016;9:88.

34. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12:357–60.

35. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44:D733–45.

36. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012;22:1760–74.

37. Harrow J, Denoeud F, Frankish A, Reymond A, Chen C-K, Chrast J, et al. GENCODE: producing a reference annotation for ENCODE. Genome Biol. 2006;7(Suppl 1):S4.1–9.

38. Liao Y, Smyth GK, Shi W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res. 2013;41:e108.

39. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30:923–30.

40. Kroll KW, Mokaram NE, Pelletier AR, Frankhouser DE, Westphal MS, Stump PA, et al. Quality control for RNA-seq (QuaCRS): an integrated quality control pipeline. Cancer Inform. 2014;13:7–14.

41. DeLuca DS, Levin JZ, Sivachenko A, Fennell T, Nazaire M-D, Williams C, et al. RNA-SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics. 2012;28:1530–2.

42. Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. Bioinformatics. 2012;28:2184–5.

43. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.

44. Kuhn RM, Haussler D, Kent WJ. The UCSC genome browser and associated tools. Brief Bioinform. 2013;14:144–61.

45. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.

46. Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. Nat Methods. 2014;11:163–6.

47. Warner JR. The economics of ribosome biosynthesis in yeast. Trends Biochem Sci. 1999;24:437–40.

48. Deng W, Aimone JB, Gage FH. New neurons and new memories: how does adult hippocampal neurogenesis affect learning and memory? Nat Rev Neurosci. 2010;11:339–50.

49. Weston NM, Sun D. The potential of stem cells in treatment of traumatic brain injury. Curr Neurol Neurosci Rep. 2018;18:1.

50. Kuruba R, Hattiangady B, Shetty AK. Hippocampal neurogenesis and neural stem cells in temporal lobe epilepsy. Epilepsy Behav. 2009;14(Suppl 1):65–73.

51. Liddelow SA, Barres BA. Reactive astrocytes: production, function, and therapeutic potential. Immunity. 2017;46:957–67.

52. Artegiani B, Lyubimova A, Muraro M, van Es JH, van Oudenaarden A, Clevers H. A single-cell RNA sequencing study reveals cellular and molecular dynamics of the hippocampal neurogenic niche. Cell Rep. 2017;21:3271–84.

53. Hochgerner H, Zeisel A, Lönnerberg P, Linnarsson S. Conserved properties of dentate gyrus neurogenesis across postnatal development revealed by single-cell RNA sequencing. Nat Neurosci. 2018;21:290–9.

54. Dulken BW, Leeman DS, Boutet SC, Hebestreit K, Brunet A. Single-cell transcriptomic analysis defines heterogeneity and transcriptional dynamics in the adult neural stem cell lineage. Cell Rep. 2017;18:777–90.

55. Shin J, Berg DA, Zhu Y, Shin JY, Song J, Bonaguidi MA, et al. Single-cell RNA-seq with waterfall reveals molecular cascades underlying adult neurogenesis. Cell Stem Cell. 2015;17:360–72.

56. Shi Z, Geng Y, Liu J, Zhang H, Zhou L, Lin Q, et al. Single-cell transcriptomics reveals gene signatures and alterations associated with aging in distinct neural stem/progenitor cell subpopulations. Protein Cell. 2018;9:351–64.

57. Zhang J, Jiao J. Molecular biomarkers for embryonic and adult neural stem cell and neurogenesis. Biomed Res Int. 2015;2015:727542.

58. Lun ATL, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol. 2016;17:75.

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

60. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods. 2014;11:740–2.

61. Ding N, Melloni L, Zhang H, Tian X, Poeppel D. Cortical tracking of hierarchical linguistic structures in connected speech. Nat Neurosci. 2016;19:158–64.

62. Tung P-Y, Blischak JD, Hsiao CJ, Knowles DA, Burnett JE, Pritchard JK, et al. Batch effects and the effective design of single-cell gene expression studies. Sci Rep. 2017;7:39921.

63. Liu H, Li Y, He J, Guan Q, Chen R, Yan H, et al. Robust transcriptional signatures for low-input RNA samples based on relative expression orderings. BMC Genomics. 2017;18:913.

64. Gertz J, Varley KE, Davis NS, Baas BJ, Goryshin IY, Vaidyanathan R, et al. Transposase mediated construction of RNA-seq libraries. Genome Res. 2012;22:134–41.

65. Slomovic S, Fremder E, Staals RHG, Pruijn GJM, Schuster G. Addition of poly(A) and poly(A)-rich tails during RNA degradation in the cytoplasm of human cells. Proc Natl Acad Sci USA. 2010;107:7407–12.

66. Ha KCH, Blencowe BJ, Morris Q. QAPA: a new method for the systematic analysis of alternative polyadenylation from RNA-seq data. Genome Biol. 2018;19:45.

Walker *et al. J Transl Med*   *(2020) 18:63*

Page 15 of 15

67. Elkon R, Ugalde AP, Agami R. Alternative cleavage and polyadenylation: extent, regulation and function. Nat Rev Genet. 2013;14:496–506.
68. Ohio Supercomputer Center. Ohio Supercomputer Center. Columbus, OH: Ohio Supercomputer Center; 1987. http://osc.edu/ark:/19495/f5s1ph73.

Publisher's Note