Evaluation of biological and technical variations in low-input RNA-Seq and single-cell RNA-Seq

Fan Gao, Jae Mun Kim, JiHong Kim and Ming-Yi Lin
Zilkha Neurogenetic Institute, Keck School of Medicine, University of Southern California, Los Angeles, CA 90089, USA
Email: fangao@broadinstitute.org
Email: jaemunki@usc.edu
Email: lomolith@gmail.com
Email: mingyili@usc.edu

Charles Y. Liu and Jonathan J. Russin
Department of Neurosurgery, Keck School of Medicine, University of Southern California, Los Angeles, CA 90089, USA
Email: cliu@usc.edu
Email: jonathan.russin@med.usc.edu

Zilkha Neurogenetic Institute, Keck School of Medicine, University of Southern California, Los Angeles, CA 90089, USA
Email: christopher.walker@med.usc.edu
Email: william.mack@med.usc.edu
Email: oleg.evgrafov@downstate.edu
Email: rchow@med.usc.edu
Email: james.knowles@downstate.edu
Email: wangk@email.chop.edu
*Corresponding authors
Abstract: Background: Low-input or single-cell RNA-Seq are widely used today, but two technical questions remain: 1) in technical replicates, what proportion of noises comes from input RNA quantity rather than variation of bioinformatics tools?; 2) In single neurons, whether variation in gene expression is attributable to biological heterogeneity or just random noise? To examine the sources of variability, we have generated RNA-Seq data from low-input (10/100/1000pg) reference RNA samples and 38 single neurons from human brains. Results: For technical replicates, the quantity of input RNA is negatively correlated with expression variation. For genes in the medium- and high-expression groups, input RNA amount explains most of the variation, whereas bioinformatic pipelines explain some variation for the low-expression group. The t-distributed stochastic neighbour embedding (t-SNE) method reveals data-inherent aggregation of low-input replicate data, and suggests heterogeneity of single pyramidal neuron transcriptome. Interestingly, expression variation in single neurons is biologically relevant. Conclusions: We found that differences in bioinformatics pipelines do not present a major source of variation.

Keywords: RNA-Seq; single-cell sequencing; bioinformatics; TopHat; RNA-Seq by expectation maximisation; RSEM; t-distributed stochastic neighbour embedding; t-SNE; principal component analysis; PCA; annotate variation; ANNOVAR; variance.


Biographical notes: Fan Gao is the supervisor of Bioinformatics Core at the Picower Institute of MIT and an associated scientist with Broad Institute of Harvard and MIT. He finished his PhD with Mayo Clinic-Mayo Graduate School in Biomedical Engineering, and was a senior research associate at the Zilkha Neurogenetic Institute and the Broad CIRM Stem Center at University of Southern California (USC).

Jae Mun Kim received his BS in Biochemistry and Physiology at the University of Washington. He received his PhD in Neuroscience at the USC, where he developed a method for in vitro transcription amplification from a single cell and analyzed next generation sequencing data. He is currently a postdoctoral scholar at the University of California, San Diego, studying age-dependent decline in neuronal plasticity in response to spinal cord injury.

JiHong Kim received his Bachelor’s in Electrical Engineering from ChungAng University, Seoul, Korea, Master’s degree in Computer Science from State University of New York at Buffalo, and PhD in Biomedical Science from the Catholic School of Medicine, Seoul, Korea. He then became a postdoctoral fellow at the USC, researching on analytical methods and pipelines for single-cell RNA-Seq data.

Ming-Yi Lin received her BS in Electrical Engineering from National Taiwan University, Taipei, Taiwan, and PhD in Neuroscience from the USC. Her research interests include understanding the mechanism and advancing treatment for neurodegenerative diseases, developing next-generation retinal prosthesis, and creating a functional and genetic mapping of the human brain.
Charles Y. Liu completed his BSE in Chemical Engineering at the University of Michigan and his PhD in Chemical/Bioengineering at Rice University. He received his MD from Yale University and his neurosurgical training at USC Affiliated Hospitals. He is currently a Professor in Neurological Surgery, Neurology, Urology, Biomedical Engineering, and Physical Therapy/Biokinesiology at USC and serves as the Director of the USC Neurorestoration Center.

Jonathan J. Russin is an Assistant Professor of Clinical Neurological Surgery and the Assistant Surgical Director of the Center for Neurorestoration at the USC. He received MD from New Jersey Medical School, had a Residency on Neurological Surgery at LAC + USC Medical Center. After residency, he completed a fellowship with a focus on cerebrovascular/skull base disease at the Barrow Neurological Institute.

Christopher P. Walker has conducted genetic research on complex genetic disorders for over 15 years. Most recently this involved NGS library preparation for single cell RNA sequencing, micro-RNA sequencing, and whole genome and exome sequencing. Through the implementation of these automated workflows, the research projects he has worked on have a seen an increase in both the efficiency and quality of the sequencing data being produced.

Reymundo Dominguez obtained PhD in Neuroscience from the USC. He subsequently had postdoctoral training at the Laboratory of Neuroendocrinology of the University of California Los Angeles, and then at the Zilkha Neurogenetic Institute of the USC.

Adrian Camarena obtained BS in Neuroscience from the USC. After graduation, he has been a laboratory technician and laboratory supervisor at the USC focusing in RNA sequencing, and is currently a MD candidate at the University of Chicago.

Joseph D. Nguyen is a laboratory technician at the USC focusing on next-generation sequencing technology development and application.

Jennifer Herstein is a Programmer/Analyst. She worked on bioinformatics of next-generation sequencing data at USC. She received a BS in Engineering from Harvey Mudd College and an MS in Electrical Engineering from USC. Before joining USC, she was a software engineer at the Jet Propulsion Laboratory, the Capital Group, and the California Institute of Technology.

William Mack is an Associate Professor of Neurosurgery at USC. He received his Bachelor degree with honors from Cornell University. He then attended Columbia University, College of Physicians and Surgeons, where he graduated from medical school and completed Neurosurgical residency training. He spent one year performing a post-doctoral cerebrovascular research fellowship at Columbia. Following residency, he completed a Neuroendovascular Surgery fellowship at UCLA.

Oleg V. Evgrafov is a Professor at the SUNY Downstate Medical Center. He received a Master’s degree from Moscow Physical Engineering Institute, and a PhD degree from the Institute of General Genetics, Moscow. Before joining SUNY, he has been a Research Scientist at Columbia University and New York State Psychiatric Institute, as well as an Assistant Professor and later Associate Professor at USC. His research focused on the development of in vitro cellular models to study complex transcriptional alterations in various psychiatric disorders.
Robert H. Chow is a Professor of Physiology and Neuroscience and of Biomedical Engineering at the Keck School of Medicine of USC. He trained in electrophysiology and membrane biophysics with Prof. Clay Armstrong at University of Pennsylvania and with Prof. Erwin Neher (inventor of the patch clamp technique) at the Max Planck Institute for Biophysical Chemistry (Goettingen, Germany). He has applied patch clamp technique to the study of fetal and adult human brain neurons, correlating the location, morphology, electrophysiology, and transcriptomes of individual identified neurons within living brain slices.

James A. Knowles is a Professor and Chairman of Cell Biology, SUNY-Downstate Medical Center, New York. He received a Bachelor’s degree from the University of Michigan, a MD/PhD from the Albert Einstein College of Medicine and performed residency at Columbia University. Before joining SUNY, he has been an Assistant Professor and Associate Professor of Psychiatry at the College of Physicians and Surgeons, Columbia University, and later as a Professor and the Della Martin Chair of Psychiatry & Neuroscience at USC. Most of his work has been on the genetics of the complex psychiatric disorders.

Kai Wang is an Associate Professor at the Raymond G. Perelman Center for Cellular and Molecular Therapeutics of the Children’s Hospital of Philadelphia (CHOP) and the Perelman School of Medicine at the University of Pennsylvania (Penn). He received a Bachelor’s degree from Peking University, a master’s degree from Mayo Clinic, and a PhD from the University of Washington. He had postdoctoral training at Penn and CHOP, then became an Assistant and later Associate Professor at USC, and then an Associate Professor at Columbia University Medical Center. His research focused on the development and application of genomic approaches to study the genetic basis of human diseases.

1 Introduction

With the development and deployment of next generation sequencing (NGS) techniques, RNA-Seq has now become a standard tool to investigate the transcriptional landscape of biological systems (Morozova et al., 2009; Wang et al., 2009). With rapidly evolving new low-input RNA sequencing library preparation methods coupled with the declining cost of the NGS, RNA-Seq on the single-cell level has become a reality in the past few years. There have been already several interesting attempts, including our own, to apply RNA-Seq to different types of single cells and identify the variation of expression in a small population of cells, such as neuronal cells (Tang et al., 2009; Islam et al., 2011; Qiu et al., 2012; Deng et al., 2014; Camp et al., 2015; Darmanis et al., 2015; Johnson et al., 2015; Zeisel et al., 2015). Considering that cells in the brain are composed of highly heterogeneous populations both spatially and temporally, single-cell transcriptomic profiling is an attractive approach to explore the molecular differences underlining the phenotypic heterogeneity. Compared to conventional RNA-Seq of bulk RNA extraction from tissue homogenates, single-cell RNA-Seq requires better control of technical variability, because of the intrinsic technical difficulties of processing the low mass of RNA from single-cells. Furthermore, the ability of bioinformatics methods to properly
handle these data and address biases becomes a critical issue to extract more biological insights from single cell studies (Hebenstreit, 2012).

In addition to biologically meaningful variation (i.e., biological noise), potential technical noise from single-cell libraries mainly come from the low amount of single cell RNA material (thus batch effect of library preparation and sequencing runs becomes more obvious). Evaluation of variability from technical replicates of low input material helps us better understand the nature of technical noise.

An additional, and often ignored, the source of variation comes from the use of different bioinformatics pipelines, as some of them may be more robust against technical noises than others, especially in the analysis of single-cell data. One of the primary uses of RNA-Seq is the estimation of gene abundance. Gene quantification tools such as TopHat/Cufflinks and RNA-Seq by expectation maximisation (RSEM) were initially developed to estimate transcript/gene expression for RNA-Seq data generated using micrograms of input RNA, using Fragments Per Kilobase of transcripts per Million mapped reads (FPKM) as a measure. Both methods have been recently applied to analyse single-cell RNA-Seq data (Shalek et al., 2013; Trapnell et al., 2014), however, it is interesting to compare their performance side-by-side using low-input technical replicate data. Reduction of high dimensionality to 2D space for visualisation is another challenging issue for single-cell RNA-Seq data, since one of the common biological questions in single-cell studies is the grouping/clustering of cells based on a whole-transcriptome expression. Linear PCA and non-linear t-distributed stochastic neighbour embedding (t-SNE) methods are often used for visual examination of data clusters (Lee et al., 2014; Macosko et al., 2015), and a comparison of their performance on technical replicates is useful to better understand the noise of our assay condition.

Lastly, single-cell RNA-Seq data generally have a higher degree of variation compared to bulk RNA-Seq data of biological replicates. Considering 10pg of total RNA roughly correspond to RNA content of single neurons, a comparison of the variation in gene expression in single-neurons to 10pg of total reference RNA input into the same RNA-Seq protocol is valuable to understand the difference of biological noises and technical noises.

In the current study, we generated a set of replicate data specifically to assess the technical noises of single-cell data and evaluate the sources of the noises (i.e., the influence of bioinformatics pipelines or quantity of starting materials upon detected variation of gene expression). Furthermore, we analysed a set of single-cell RNA-Seq data from neurons collected from human brain slices, to assess the presence of biological noises and evaluate how best to capture such noises in the presence of other sources of noises. We performed saturation analysis, variation analysis and dimensionality reduction analysis on these datasets to understand the intrinsic variations/noises and optimise an analysis pipeline.

2 Material and methods

2.1 Human low-input RNA-Seq datasets

Two sets of human reference RNAs were used in our comparative analysis on low-input RNA-Seq, with the purpose of creating technical replicate data for analysis of noises. We used Human Brain Reference [ThermoFisher Scientific, Grand Island, NY] with each
quadruplicate of 10 pg and 100 pg and one triplicate of 1000 pg, and Universal Human Reference [Agilent Technologies, Santa Clara, CA] with 6 replicates of 10 pg and triplicates for each 100 pg and 1000 pg. The HBR set is a pool of RNA from several brain regions, with a comprehensive representation of brain-expressed genes, yet the UHR was generated by pooling RNA from 10 human cell lines sourced from different tissues, so it has a broad representation from genes that can be expressed in human tissues. We performed cDNA and library preparation using the NuGen Ovation® RNA-Seq System V2 followed by Ovation Rapid Library system (NuGen, San Carlos, CA) with slight modifications, and sequenced them on the Illumina HiSeq 2500 sequencer with single-end 101 bp reads using Rapid v1 flow cells and sequencing kit. On average, we generated 21,114,302 reads per sample.

2.2 Human single-neuron RNA-Seq datasets

Brain tissue was donated with informed consent and procurement procedure was approved by an institutional review board of USC. Specimens were transported in Hypothermosol solution (Sigma, USA; Biolife Solutions, USA) on ice. Tissue transport time was always kept under half an hour. Brain tissue was then visually examined for structural integrity and the anterior or posterior face of the tissue was glued onto specimen plate. The tissue was sliced into 400 µm slices in cold (4°C), oxygenated (95%O₂, 5% CO₂) artificial cerebral spinal fluid (ACSF; NaCl 124 mM; KCl 4 mM; NaHCO₃ 26 mM; Glucose 10 mM; CaCl₂ 2 mM; MgCl₂ 2 mM) using a vibratome (Leica VT1200S, Germany). Vertical deflection of the blade was minimised with Vibrocheck technology and slicing parameters were: speed 0.1–0.14 mm/s and vibration amplitude 1.5–2 mm. Slices were transferred to a recovery chamber (32°C, 95%O₂, 5% CO₂) and allowed to recover in ACSF for 30–60 min before recording and cell collection was performed.

Neurons were visually identified under illumination with an infrared Dodt gradient contrast system. Patch pipettes (6–10 MOhm; 1.2 mm O.D.) were filled with intracellular solution (K-gluconate 130 mM; KCl 2 mM; CaCl₂ 1 mM; MgATP 4 mM; GTP 0.3 mM; phosphocreatine 8 mM; HEPES 10 mM; EGTA 11 mM; pH 7.25 and 300 mOsm) containing RNase inhibitor (0.4 U/µl; RNase inhibitor, Clontech). Pressure control of the patch pipette was performed with an automatic pressure control unit (ez-gSEAL 100b, Neobiosystem, USA). Before reaching the cell, a positive pressure (25–50 mmHg) was applied, and once the pipette touched the cell membrane, a gentle suction (−15 to −30 mmHg) was applied to form giga-seal. The seal was allowed to stabilise for 1 min and the membrane rupture for whole-cell was done by applying a brief pulse of suction (−90 to −150 mmHg for 400 ms). The RNA content from somatic cytoplasm of the cell was collected into the patch pipette by applying a continuous strong negative pressure (−200 mmHg to −250 mmHg) for at least 5 min or until the cell was completely collected into the patch pipette. The content was expelled into a PCR tube containing 5 µl of lysis buffer (5g/L NaCl; 1% Triton X-100; 1% NP-40; 5% sodium deoxycholate; 5% Tris-HCl; 20mM HEPES) by breaking the pipette tip and applying positive pressure (25–50 mmHg). The samples were flash frozen in liquid nitrogen before storing in −80°C.

The right temporal pole region (middle gyrus) of two adult brain samples was targeted, with 16 and 22 single cells collected per sample. On average, 5,780,246 reads per sample were generated. Prior to analysis, we remove adaptor sequences and low-
Evaluation of biological and technical variations

quality reads ($Q < 25$ or length $< 20$bp) by cutadapt (http://cutadapt.readthedocs.org/en/stable/).

2.3 Quantification and clustering analysis on RNA-Seq data

To assess the noises introduced by the use of different bioinformatics approaches, we processed the RNA-Seq data using two widely used bioinformatics pipelines, TopHat/CuffLinks (Trapnell et al., 2012) and RSEM (Li and Dewey, 2011). The default parameters for each pipeline were used in data processing, by aligning the raw reads to the hg19 genome assembly with the UCSC hg19 transcript models. Gene expression was estimated using Fragments Per Kilobase of transcripts per Million mapped reads (FPKM). Pairwise Pearson’s correlation coefficients were calculated to construct a correlation matrix for hierarchical clustering. Furthermore, to perform dimensionality reduction of RNA-Seq data for visualising transcriptome similarities, we applied both principal component analysis (PCA) and t-SNE to generate 2D maps, using selected genes with mean expression above a threshold value.

2.4 Data variation analysis

Squared coefficient of variation ($CV^2$) value was used as a metric for evaluation of noise for different conditions. To calculate relative proportion of variance (relative sum of squares) for each gene, ANOVA test was carried out for two factors: RNA input amounts and bioinformatics pipelines used for the analysis, using custom R codes (www.R-project.org).

2.5 Other analyses

A permutation-based $R$ script was used to calculate statistics for the $CV^2$ value of each gene. The false-discovery rate (FDR) was calculated based on 10,000 times of random permutations. Genes with FDR < 0.05 were used for web-based TOPPGENE gene ontology analysis (Chen et al., 2009). The top 10 ‘molecular function’ and ‘biological process’ terms ranked by Bonferroni corrected q-values were presented, and the ontology terms with $q < 0.05$ were considered as significantly enriched. All data used in the study are available from the Single Cell Analysis Program – Transcriptome Project (http://scap-t.org/).

3 Results

3.1 Saturation of low-input RNA-Seq data

To establish a reference dataset for low-input/single cell RNA-Seq, two human reference RNA samples — Human Brain Reference (HBR) and Universal Human Reference (UHR), were sequenced with 3 different RNA quantities (10pg, 100pg and 1000pg), each with 3–6 replicates. To evaluate the consistency of data complexity from sequenced replicate libraries, percentages of unique reads were calculated, and two libraries with low complexity (less than 50% of unique reads) were removed from further analysis (Supplementary Figure 1). The percentage of unique reads in the remaining samples averaged at ~80%.
To understand how sequencing depth and input RNA amount affect the number of genes detected, saturation curves were constructed by subsampling the raw sequencing reads from each library and determining the number of genes detected at different sequencing depth. Detection of the genes with high expression (FPKM > 1) saturated around 1 million reads, while the number of genes detected with FPKM > 0.1 barely approached saturation at approximately 5 million reads (Figure 1). Therefore, in our assay conditions, 1 million sequencing reads provide sufficient sequencing depth to detect highly expressed genes from low input RNA samples. Importantly, when input RNA amount decreases from 100pg to 10pg, the number of genes detected decreases drastically (more than 20% for UHR, more than 30% for HBR), suggesting the likelihood of insufficient process of converting RNA molecules into sequencing libraries at 10pg, as well as the possibility that lowly transcribed molecules are not sampled from such low input material.
3.2 Technical noise analysis

It is expected that datasets from single-cell RNA-Seq contain more technical noise than conventional RNA-Seq because of the challenges in processing the minute amount (roughly 10pg) of starting material from an individual cell; however, the extent to which RNA quantity affects technical noises is largely unknown. Furthermore, it is unclear whether some bioinformatics algorithms and software tools are more resilient to technical noises than others in the context of single-cell RNA-Seq.

We first evaluated the performance of two popular gene quantification tools TopHat/Cufflinks and RSEM in generating a variation of expression values (FPKM) from technical replicates. For the HBR/UBR human RNA-Seq data, the calculated gene CV² (squared coefficient of variation) values using either tool are largely similar to 1000pg and 100pg input amount (Figure 2). However, at a very low input level that resembles single-cell data (10pg), a portion of the genes with low CV², as determined by RSEM, have increased CV² values by the TopHat/Cufflinks pipeline (Figure 2).

Genes with high expression levels are expected to generate more consistent output from technical replicates. As shown in Figure 3, for HBR sample, with 1000pg and 100pg input RNA, CV² is generally negatively correlated with gene expression levels, consistent with our expectation. However, with 10pg input RNA, this relationship becomes obscure, suggesting that factors other than quantified expression value may affect CV² of HBR; however, for UHR sample, 100pg input shows a better correlation (negative) than the 1000pg input. These observations are consistent with either of the two computational pipelines (RSEM, TopHat/Cufflinks).
Figure 3  Technical variance of gene expression as a function of expression levels. The CV² value and the mean FPKM value for each gene were used for the scatter plots. Six different sample conditions subject to either TopHat/Cufflinks or RSEM bioinformatics pipelines were analysed (see online version for colours)
3.3 Noise contributed from input RNA amount and bioinformatics pipeline

In our assay design, three different input RNA amounts and two different bioinformatics pipelines were tested to evaluate noise/variation from technical replicates. Three equal-sized gene groups were generated (low, medium and high) based on mean FPKM values. We further carried out ANOVA test to evaluate how CV² is influenced by amounts of input RNA and bioinformatics pipelines. Calculated percentage of variation explained by either input RNA amount or bioinformatics pipeline was shown in Figure 4. As noted, for genes in the medium and high expression groups, most of the variation for both HBR and UHR samples can be explained by input RNA (for 75% of genes, more than 90% of CV² explained by input RNA), while the bioinformatics pipeline has minimal effect. However, for genes in the low expression group, input RNA has less contribution to variation (only half of the genes have more than 90% of CV² explained by input RNA), while bioinformatics pipeline appears to explain more of variation, but even taken together these factors do not explain all variations in gene expressions. Therefore, for lowly-expressed genes, factors other than input RNA amount or bioinformatics pipeline may contribute to technical variation.

Figure 4 ANOVA estimates of percentages of the variance of gene expression (CV²) contributed to either input RNA amount or bioinformatics pipeline. The percentage values were calculated for each gene, and presented as boxplots for all three defined groups of genes (low-, medium- and high-expression groups)

3.4 Data clustering and dimensionality reduction

In our previous variation analysis, we found that the two bioinformatics pipelines, RSEM and TopHat/Cufflinks, are largely consistent. We decided to choose RSEM processed data for the analysis below because of the computational speed of RSEM. We first calculated Pearson’s correlation coefficients between different replicate data of human HBR/UHR samples based on FPKM values of the expressed genes (genes with mean FPKM greater than 0.1 or 1), and plotted the results as colour-coded correlation matrix.
ordered by hierarchical clustering (Figure 5). Compared to technical replicate data generated from higher amounts of input RNA (100pg and 1000pg), replicates from the 10pg input are less consistent with lower correlation values. Also noted, in spite of different FPKM cutoff to use, hierarchical clustering based on correlation of the whole transcriptome clearly separates the HBR replicate data from the UHR replicate data. Therefore, low-input transcriptome profiling is capable of generating biologically meaningful and sample-specific transcriptomic signatures that overcome technical noise.

**Figure 5** Correlation heatmaps of the low-input transcriptomic data. Colour-scale represents Pearson’s correlation coefficient. Genes with mean FPKM values greater than 1 or 0.1 were selected for the analysis (see online version for colours)

To further visualise multidimensional gene expression data, we performed both PCA and t-distributed stochastic neighbour embedding (t-SNE) analysis. In contrast to the standard linear PCA method, t-SNE method can capture nonlinear relationships in data (Jamieson et al., 2010). As shown in Figure 6, PCA analysis clearly separated HBR and
UHR libraries with the first two dimensions; however, the replicate libraries were distributed sparsely in the 2D space, despite the FPKM cutoff applied. In contrast, t-SNE generally generated better clustering results than PCA method. Thus, t-SNE visualisation reduced the non-linear variations associated with low-input RNA-Seq data.

**Figure 6** 2D plots presenting the sample clustering results from two different dimensionality reduction algorithms (PCA and t-SNE). A total of 21 low-input transcriptomic data were shown as colour-coded dots in the plots (see online version for colours)

### 3.5 Single neuron analysis

After evaluating technical noises from different aliquots of the same RNA samples, we assayed biological noises from single neuronal cells. We generated 38 single-cell RNA-Seq datasets on single temporal pyramidal neurons from two adults who underwent brain surgery. Similar to our previous analysis on HBR/UHR data, saturation analysis revealed that 1 million reads are sufficient to detect most of the expressed genes that were converted into libraries (Figure 7(A)). For the two individuals, an average of 5000–6000 known genes can be detected from single neurons. By using publicly available tissue RNA-Seq data generated from the middle frontal gyrus region of an adult brain (GEO record: GSM1173807), our gene quantification analysis detected ~13,500 known genes. Thus, our single neuron data can cover roughly 40% of the tissue transcriptome (a heterogeneous cell population, including neurons, astrocytes, microglia, oligodendrocytes, and other cell types), and provide a high spatial resolution map of neuron-specific transcriptome.

Unlike 10pg low-input replicate data on the HBR brain reference sample, single pyramidal neuron data showed much greater variation (Figure 7(B)) of FPKM values for genes. A low Pearson’s correlation coefficient ($r = 0.366$) of gene CV$^2$ values between HBR and single neuron data revealed that variations of gene expressions in single neurons are different from those in HBR, suggesting an intrinsic biological variation of gene expression at single cell level, other than technical noise observed for low-input
data. Unlike aggregated clustering pattern observed for low-input technical replicate data (Figure 6), t-SNE 2D map showed sub-clustering of single neuron transcriptome (Figure 7(C)), suggesting cellular heterogeneity.

Figure 7 Analysis of single-cell data. (A) Saturation analysis of number of genes detected as a function of number of sequencing reads; (B) variance values (CV²) of gene expression compared between a single-cell data and low-input HBR data; (C), t-SNE 2D plot of single-cell data collected from two brains (AA32 and AA35) and (D) enrichment gene ontology terms for the most variably expressed genes in the single cell data (see online version for colours).

Indeed, gene ontology analysis (Figure 7(D)) of the top CV² genes (1449 genes with permutation-based FDR < 5%, see Materials and Methods for details) in single neuron data showed sensory perception and stimulus detection related terms as the most enriched terms in “Biological Process”. In contrast, the gene ontology analysis for the most variable genes in HBR 10pg data failed to identify any statistically significant “Biological Process” terms (Supplementary Figure 2). Pyramidal neurons are excitatory neurons in the brain that mediates neuroplasticity and cognition. Highly variable expression of “sensory perception and stimulus detection” genes from individual pyramidal neurons may represent functional heterogeneity in processing different sensory modalities/inputs.
4 Discussion

In this study, we generated a set of reference RNA-Seq datasets on different quantities of RNA samples, and analysed the technical noises. We found that variation contributed from bioinformatics pipeline is generally minor compared to the quantity of input RNA. We also demonstrated that t-SNE is more effective than PCA to handle the noises from very low-input RNA-Seq (single-neuron level).

One common concern for single-cell analysis is the increased proportion of duplicate reads from sequencing runs. In our study, saturation analysis for both low-input and single-neuron data showed almost saturated gene detection with merely 1 million sequencing reads, suggesting that increasing the multiplexity of single-cell library (e.g., 96 single cells for one HiSeq2000 lane) is more meaningful (also cost-effective) than chasing sequencing depth of individual single cells.

One potential limitation of our study is that only one library preparation protocol is used in our study, so it may not extrapolate well to other single-cell sequencing approaches. Currently, there are several different commercially available single-cell library prep kits on the market. In our assay condition (NuGen kit), we observed an average of 27% library complexity for single neuron data. In our experience, this measure is very similar to our previous results generated from the SMART-Seq kit.

Another limitation of our study is that we only compared two bioinformatics pipelines, namely TopHat/Cufflinks and RSEM. It is unclear whether other software tools may be more robust to single-cell analysis, especially with respect to the quantification of gene expression levels. Furthermore, several imputation methods have been developed to ‘predict’ the gene expression levels of certain genes based on expression of other genes, to address the allelic dropout problem. We did not implement these approaches in our analysis, though they may reduce the noises observed in a single-cell data. Also, bioinformatic detection of splicing and structural variation will be very interesting for understanding the biology of single neurons.

Variations/noises for single-cell RNA-Seq data are considered to be contributed to biological noises (cellular heterogeneity), technical noises and bioinformatics noises. In our study, we demonstrated some clear examples that t-SNE can reduce technical noises and capture intrinsic biological noises for visualising biologically meaningful sub-grouping of single cells. Finally, we wish to stress that the RNA-Seq data that we have generated as part of the study will be an invaluable resource for the community, to benchmark different experimental and bioinformatics approaches to generate or analyse single-cell RNA-Seq data.

Acknowledgements

This study was supported by NIH grant U01 MH098937 (J.A.K. and R.C.) and R01 HG006465 (K.W.). We thank our fellow members of the Single Cell Analysis Program – Transcriptome project (SCAP-T), for helpful discussions and comments.
Author contributions


References


Evaluation of biological and technical variations


**Supplementary information**

**Figure 1**  Library complexity of low-input RNA-Seq data (see online version for colours)
Figure 2  Enriched gene ontology terms for the most variably expressed genes in low-input RNA-Seq data

<table>
<thead>
<tr>
<th>Molecular Function</th>
<th>Name</th>
<th>q-value Bonferroni</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0004844</td>
<td>uracil DNA N-glycosylase activity</td>
<td>1.42E-01</td>
</tr>
<tr>
<td>GO:0097506</td>
<td>deaminated base DNA N-glycosylase activity</td>
<td>1.42E-01</td>
</tr>
<tr>
<td>GO:0019104</td>
<td>DNA N-glycosylase activity</td>
<td>1.58E-01</td>
</tr>
<tr>
<td>GO:0016796</td>
<td>hydrolase activity, hydrolyzing N-glycosyl compounds</td>
<td>2.40E-01</td>
</tr>
<tr>
<td>GO:0000701</td>
<td>purine-specific mismatch base pair DNA N-glycosylase activity</td>
<td>2.61E-01</td>
</tr>
<tr>
<td>GO:0000700</td>
<td>mismatch base pair DNA N-glycosylase activity</td>
<td>5.95E-01</td>
</tr>
<tr>
<td>GO:0017088</td>
<td>single-strand selective uracil DNA N-glycosylase activity</td>
<td>1.00E+00</td>
</tr>
<tr>
<td>GO:0073894</td>
<td>deoxyribonucleoside 5-monophosphate N-glycosidase activity</td>
<td>1.00E+00</td>
</tr>
<tr>
<td>GO:0070694</td>
<td>nicotinamide riboside hydrolase activity</td>
<td>1.00E+00</td>
</tr>
<tr>
<td>GO:0070634</td>
<td>nicotinic acid riboside hydrolase activity</td>
<td>1.00E+00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Biological Process</th>
<th>Name</th>
<th>q-value Bonferroni</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0019386</td>
<td>galactose catabolic process</td>
<td>1.00E+00</td>
</tr>
<tr>
<td>GO:0032301</td>
<td>regulation of iocosanoid secretion</td>
<td>1.00E+00</td>
</tr>
<tr>
<td>GO:0017004</td>
<td>cytochrome complex assembly</td>
<td>1.00E+00</td>
</tr>
<tr>
<td>GO:0071719</td>
<td>iocosanoid transport</td>
<td>1.00E+00</td>
</tr>
<tr>
<td>GO:1901671</td>
<td>fatty acid derivative transport</td>
<td>1.00E+00</td>
</tr>
<tr>
<td>GO:0066626</td>
<td>regulation of canonical Wnt signaling pathway</td>
<td>1.00E+00</td>
</tr>
<tr>
<td>GO:0048641</td>
<td>regulation of skeletal muscle tissue development</td>
<td>1.00E+00</td>
</tr>
<tr>
<td>GO:0051729</td>
<td>protein de-ADP-ribosylation</td>
<td>1.00E+00</td>
</tr>
<tr>
<td>GO:0035903</td>
<td>corticosteroid hormone secretion</td>
<td>1.00E+00</td>
</tr>
<tr>
<td>GO:1901601</td>
<td>quinone metabolic process</td>
<td>1.00E+00</td>
</tr>
</tbody>
</table>