zUMIs
A fast and flexible pipeline to process RNA sequencing data with UMIs

Swati Parekh[1,2*], Christoph Ziegenhain[1*], Beate Vieth[1], Wolfgang Enard[1], Ines Hellmann[1,2]

[1] Anthropology & Human Genomics, Department of Biology II, Ludwig-Maximilians University, Munich, Germany
[2] corresponding author

*** contributed equally**

Abstract

RNA sequencing is increasingly performed with less starting material and at a higher sample throughput, e.g. to analyse single-cell transcriptomes. In this context, unique molecular identifiers (UMIs) are used to reduce amplification noise and sample-specific barcodes are used to track libraries. Here, we present a fast and flexible pipeline to process data from such RNA-seq protocols.
Availability: https://github.com/sdparekh/zUMIs

1 Introduction

The recent development of sensitive protocols allows to generate RNA-seq libraries of single cells [1]. The throughput of such scRNA-seq protocols is rapidly increasing, enabling the profiling of tens of thousands of cells [2, 3] and opening exciting possibilities to analyse cellular identities [4, 5]. As the required amplification from such low starting amounts introduces substantial amounts of noise [6], many scRNA-seq protocols incorporate unique molecular identifiers (UMIs) to label individual cDNA molecules with a random nucleotide sequence before amplification [7]. This allows to computationally remove amplification noise and thus increases the power to detect expression differences [8, 9]. To increase the throughput, many protocols also incorporate sample-specific barcodes (BCs) to label all cDNA molecules of a single cell with a nucleotide sequence before library generation [10, 2]. Additionally, for cell types such as neurons it has

Name	Reference	Open Source	Quality UMI/BC	Mapper	intron counting	Down-sampling
CellRanger	[2]	no	no	STAR	no	yes
Drop-seq	[10]	no	yes	STAR	no	no
CEL-seq	[13]	yes	yes	bowtie2	no	no
umis	[14]	yes	no	Kallisto	no	no
zUMIs	This work	yes	yes	STAR	yes	yes

Table 1. Pipelines handling UMI expression data

proven to be more feasible to isolate RNA from single nuclei rather than whole cells [11, 12]. This decreases mRNA amounts further, so that it has been suggested to count intron-mapping reads as part of nascent RNAs. However, the few bioinformatic tools that process RNA-seq data with UMIs and BCs have limitations with respect to availability, mapping, quality assessment and/or can not consider intronic reads (Table 1). Here, we present *zUMIs*, a fast and flexible pipeline to overcome such limitations.

2 zUMIs

zUMIs is a pipeline that processes paired fastq files containing the UMI and BC in one read and the cDNA sequence in the other read, filters out reads with bad BCs or UMIs based on sequence quality, maps reads to the genome and outputs count tables of unique UMIs or reads per gene (Figure 1). To allow the quantification of intronic reads that are generated from unspliced RNAs especially when using nuclei as input material, three separate count tables for exons, introns and exon+introns are provided. Another unique feature of *zUMI* is that it allows for downsampling of reads before summarizing UMIs per feature, which is recommended for cases of highly different read numbers per sample [15]. *zUMIs* is flexible with respect to the length and sequences of the BC and UMIs, making it compatible with a large number of protocols [16, 17, 10, 13, 3, 2].

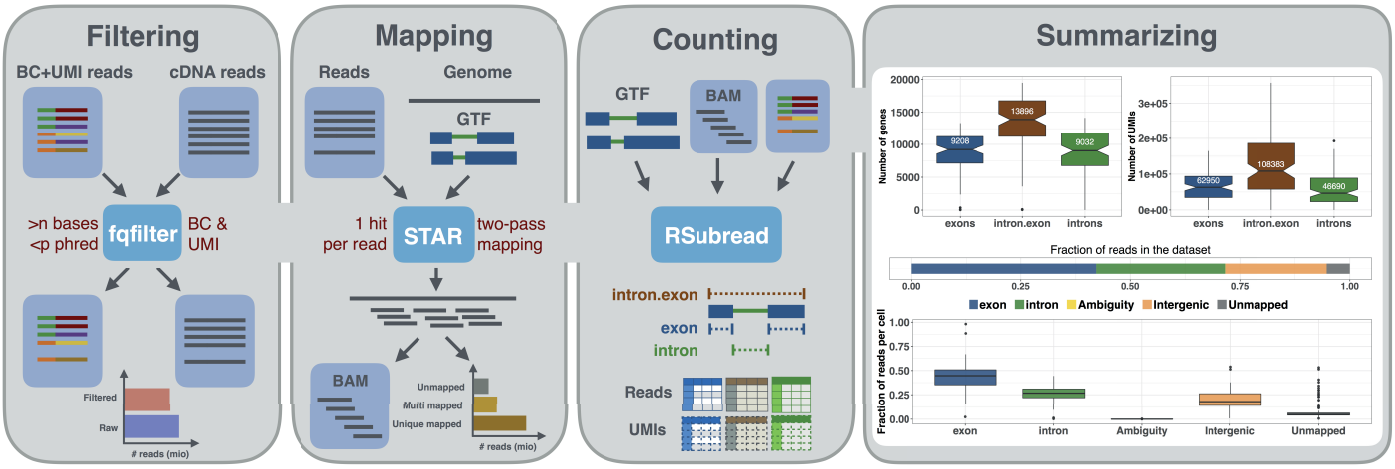

Figure 1. zUMIs schematic overview.

A Each of the grey panels from left to right depicts a step of the *zUMIs* pipeline. First, paired fastq files are filtered according to user-defined BC and UMI quality thresholds. Next, the remaining cDNA reads are mapped to the reference genome using STAR. Then gene-wise read and UMI count tables are generated for exon, intron and exon+intron overlapping reads. To obtain comparable library sizes, reads can be downsampled to a desired range during the counting step. Optionally, *zUMIs* also generates data and plots for several quality measures, such as the number of detected Genes/UMIs per barcode and distribution of reads into mapping feature categories (Supplementary Figure 3).

2.1 Processing pipeline

The input for *zUMIs* is a pair of fastq files, whereas one file contains the cDNA sequences and the other one the read containing the BC and UMI. The exact location and length of UMI and BC are specified by the user. Note that both fastq files need to be ordered by read name, which is usually the case if unprocessed files are used. The first step in our pipeline is to filter reads where the BC or the UMI fails a user-defined quality threshold. This helps to eliminate spurious BCs and is expected to reduce noise. The cleaned-up reads are then mapped to the genome using the splice-aware aligner STAR [18]. The user is free to adapt the STAR options to their data, however *zUMIs* requires that only one mapping position per read is reported. Next, reads are assigned to genes and to exons or introns based on the provided gtf file, whereas introns are defined as not overlapping with any exon. Rsubread featureCounts [19] is used to first assign reads to exons and afterwards to check whether the remaining reads fall into introns. The resulting output is then read into R using data.table [20] and count tables for UMIs and reads are generated. *zUMIs* tabulates the UMIs/gene either for user-specified BCs or for the n BCs with the highest read counts.

2.2 Output and statistics

zUMIs outputs three UMI and three read count tables: one for traditional exon mapping gene-wise counts, one for intron and one for intron+exon counts. If a user chooses the downsampling option, 6 additional count-tables are provided in which samples with an excess of reads are downsampled and samples with too few reads are dismissed (Supplementary Figures 4). We highly recommend to use this option, because normalizing across samples with vastly different library sizes does not work well [15, 21]. *zUMIs* also reports descriptive statistics. To evaluate library quality *zUMIs* summarizes the fractions of unmapped, ambiguously mapped, exon and intron mapped reads and to evaluate library complexity, the numbers of detected genes and UMIs per sample are provided (Supplementary Figures 2,3).

We processed 227 million reads with *zUMIs* and quantified expression levels for exonic and intronic counts on a unix machine using up to 16 threads, which took barely 3 hours. Increasing the number of reads increases the processing time approximately linearly, whereas filtering, mapping and counting each take up roughly one third of the total time (Supplementary Figure 1).

3 Conclusions

zUMIs is a fast and flexible pipeline to process raw reads to count tables for RNA-seq data using UMIs. To our knowledge it is the only open source pipeline that has a barcode and UMI quality filter, allows intron counting and has an integrated downsampling function (Table 1). These features ensure that *zUMIs* is applicable for most experimental designs of RNA-seq data, such as single-nuclei sequencing techniques [11, 12, 22], droplet based methods where the BC is unknown and the library sizes can vary a lot as well as plate-based UMI-methods with known BCs.

Funding

This work has been supported by the DFG through SFB1243 subprojects A14/A15.

Availability

The pipeline is freely available at https://github.com/sdparekh/zUMIs.

References

1. Rickard Sandberg. Entering the era of single-cell transcriptomics in biology and medicine. *Nat. Methods*, 11(1):22–24, January 2014.

2. Grace X Y Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, Mark T Gregory, Joe Shuga, Luz Montesclaros, Jason G Underwood, Donald A Masquelier, Stefanie Y Nishimura, Michael Schnall-Levin, Paul W Wyatt, Christopher M Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D Ness, Lan W Beppu, H Joachim Deeg, Christopher McFarland, Keith R Loeb, William J Valente, Nolan G Ericson, Emily A Stevens, Jerald P Radich, Tarjei S Mikkelsen, Benjamin J Hindson, and Jason H Bielas. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, 8:14049, 16 January 2017.

3. Alexander B Rosenberg, Charles Roco, Richard A Muscat, Anna Kuchina, Sumit Mukherjee, Wei Chen, David J Peeler, Zizhen Yao, Bosiljka Tasic, Drew L Sellers, Suzie H Pun, and Georg Seelig. Scaling single cell transcriptomics through split pool barcoding. 2 February 2017.

4. Allon Wagner, Aviv Regev, and Nir Yosef. Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.*, 34(11):1145–1160, 8 November 2016.

5. Aviv Regev, Sarah Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, Hans Clevers, Bart Deplancke, Ian Dunham, James Eberwine, Roland Eils, Wolfgang Enard, Andrew Farmer, Lars Fugger, Berthold Gottgens, Nir Hacohen, Muzlifah Haniffa, Martin Hemberg, Seung K Kim, Paul Klenerman, Arnold Kriegstein, Ed Lein, Sten Linnarsson, Joakim Lundeberg, Partha Majumder, John Marioni, Miriam Merad, Musa Mhlanga, Martijn Nawijn, Mihai Netea, Garry Nolan, Dana Pe'er, Anthony Philipakis, Chris P Ponting, Stephen R Quake, Wolf Reik, Orit Rozenblatt-Rosen, Joshua R Sanes, Rahul Satija, Ton Shumacher, Alex K Shalek, Ehud Shapiro, Padmanee Sharma, Jay Shin, Oliver Stegle, Michael Stratton, Michael J T Stubbington, Alexander van Oudenaarden, Allon Wagner, Fiona M Watt, Jonathan S Weissman, Barbara Wold, Ramnik J Xavier, Nir Yosef, and Human Cell Atlas. The human cell atlas. 8 May 2017.

6. Swati Parekh, Christoph Ziegenhain, Beate Vieth, Wolfgang Enard, and Ines Hellmann. The impact of amplification on differential expression analyses by RNA-seq. *Sci. Rep.*, 6:25533, 9 May 2016.

7. Teemu Kivioja, Anna Vähärautio, Kasper Karlsson, Martin Bonke, Martin Enge, Sten Linnarsson, and Jussi Taipale. Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods*, 9(1):72–74, January 2012.

8. Christoph Ziegenhain, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann, and Wolfgang Enard. Comparative analysis of Single-Cell RNA sequencing methods. *Mol. Cell*, 65(4):631–643.e4, 16 February 2017.

9. Beate Vieth, Christoph Ziegenhain, Swati Parekh, Wolfgang Enard, and Ines Hellmann. powsimr: Power analysis for bulk and single cell RNA-seq experiments. 15 March 2017.

10. Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, John J Trombetta, David A Weitz, Joshua R Sanes, Alex K Shalek, Aviv Regev, and Steven A McCarroll. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 21 May 2015.

11. Blue B Lake, Rizi Ai, Gwendolyn E Kaeser, Neeraj S Salathia, Yun C Yung, Rui Liu, Andre Wildberg, Derek Gao, Ho-Lim Fung, Song Chen, Raakhee Vijayaraghavan, Julian Wong, Allison Chen, Xiaoyan Sheng, Fiona Kaper, Richard Shen, Mostafa Ronaghi, Jian-Bing Fan, Wei Wang, Jerold Chun, and Kun Zhang. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science*, 352(6293):1586–1590, 24 June 2016.

12. Naomi Habib, Anindita Basu, Inbal Avraham-Davidi, Tyler Burks, Sourav R Choudhury, Francois Aguet, Ellen Gelfand, Kristin Ardlie, David A Weitz, Orit Rozenblatt-Rosen, Feng Zhang, and Aviv Regev. DroNc-Seq: Deciphering cell types in human archived brain tissues by massively-parallel single nucleus RNA-seq. 9 March 2017.

13. Tamar Hashimshony, Naftalie Senderovich, Gal Avital, Agnes Klochendler, Yaron de Leeuw, Leon Anavy, Dave Gennert, Shuqiang Li, Kenneth J Livak, Orit Rozenblatt-Rosen, Yuval Dor, Aviv Regev, and Itai Yanai. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.*, 17(1): 77, 28 April 2016.

14. Valentine Svensson, Kedar Nath Natarajan, Lam-Ha Ly, Ricardo J Miragaia, Charlotte Labalette, Iain C Macaulay, Ana Cvejic, and Sarah A Teichmann. Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods*, 6 March 2017.

15. Dominic Grün and Alexander van Oudenaarden. Design and analysis of Single-Cell sequencing experiments. *Cell*, 163(4):799–810, 5 November 2015.

16. Magali Soumillon, Davide Cacchiarelli, Stefan Semrau, Alexander van Oudenaarden, and Tarjei S Mikkelsen. Characterization of directed differentiation by high-throughput single-cell RNA-Seq. *bioRxiv*, 5 March 2014.

17. Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Naama Elefant, Franziska Paul, Irina Zaretsky, Alexander Mildner, Nadav Cohen, Steffen Jung, Amos Tanay, and Ido Amit. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science*, 343 (6172):776–779, 14 February 2014.

18. Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 1 January 2013.

19. Yang Liao, Gordon K Smyth, and Wei Shi. featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, 1 April 2014.

20. Matt Dowle and Arun Srinivasan. *data.table: Extension of 'data.frame'*, 2017. URL `https://CRAN.R-project.org/package=data.table`. R package version 1.10.4.

21. Ciaran Evans, Johanna Hardin, and Daniel M Stoebel. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief. Bioinform.*, 27 February 2017.

22. Suguna Rani Krishnaswami, Rashel V Grindberg, Mark Novotny, Pratap Venepally, Benjamin Lacar, Kunal Bhutani, Sara B Linker, Son Pham, Jennifer A Erwin, Jeremy A Miller, Rebecca Hodge, James K McCarthy, Martin Kelder, Jamison McCorrison, Brian D Aevermann, Francisco Diez Fuertes, Richard H Scheuermann, Jun Lee, Ed S Lein, Nicholas Schork, Michael J McConnell, Fred H Gage, and Roger S Lasken. Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons. *Nat. Protoc.*, 11(3):499–524, March 2016.