Bioinformatics, 35(21), 2019, 4472–4473
doi: 10.1093/bioinformatics/btz279
Advance Access Publication Date: 9 May 2019
Applications Note

Gene expression

The barcode, UMI, set format and BUStools

Páll Melsted [1,*], Vasilis Ntranos[2,*] and Lior Pachter[2,3,*]

[1]Faculty of Industrial Engineering, Mechanical Engineering and Computer Science, University of Iceland, Reykjavik, Iceland, [2]Division of Biology and Biological Engineering and [3]Department of Computing & Mathematical Sciences, California Institute of Technology, Pasadena, CA, USA

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on November 26, 2018; revised on February 15, 2019; editorial decision on April 11, 2019; accepted on April 13, 2019

Abstract

Summary: We introduce the **B**arcode-**U**MI-**S**et format (BUS) for representing pseudoalignments of reads from single-cell RNA-seq experiments. The format can be used with all single-cell RNA-seq technologies, and we show that BUS files can be efficiently generated. BUStools is a suite of tools for working with BUS files and facilitates rapid quantification and analysis of single-cell RNA-seq data. The BUS format therefore makes possible the development of modular, technology-specific and robust workflows for single-cell RNA-seq analysis.

Availability and implementation: http://BUStools.github.io/ and http://pachterlab.github.io/kallisto/singlecell.html.

Contact: pmelsted@gmail.com or ntranos@caltech.edu or lpachter@caltech.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The analysis of single-cell RNA-seq (scRNA-seq) data begins with three related computational tasks: read assignment to transcripts, cell determination and molecule identification (Chen *et al.*, 2018). These tasks are accomplished by utilizing information encoded in reads produced from scRNA-seq experiments. While the exact nature of the encoded information is technology specific, the components are universal: 'cell barcodes' are short sequences that identify the cells of origin for each read, 'unique molecular identifiers' (UMIs) are sequences that identify the molecule of origin for each read, and finally transcripts of origin are encoded via reverse transcribed cDNA sequences. Read assignment is typically accomplished by alignment of reads to a reference genome or transcriptome. Cell determination, which is the process of determining the valid cells in an experiment along with the reads associated to those cells, involves grouping of similar barcodes that appear to differ only due to sequencing or synthesis error, collation of the reads associated with those barcodes and determination of valid cells according to alignment statistics of the reads associated to them. Molecule identification, which is the process of determining which reads originated from the same molecule, consists of collapsing read counts when UMI sequences match in reads that have been assigned to a single transcript or gene from one cell. The challenges that must be overcome to efficiently and accurately solve the assignment, determination and identification problems are both computational and algorithmic. On the one hand, increasing throughput has resulted in large numbers of reads that make it difficult to process experiments (Svensson *et al.*, 2018). At the same time, problems of cell determination and molecule identification can benefit from innovative algorithmic ideas, e.g. (DePasquale *et al.*, 2018). Current scRNA-seq workflows confound these challenges and this has led to numerous drawbacks: software packages are frequently technology specific, the replacement of individual steps when better methods become available can be difficult, and hardware requirements may limit the scale of experiments that can be analyzed.

We introduce a new file format for representing scRNA-seq data called BUS that is an abbreviation for Barcode, UMI and Set. BUS format consists of a binary representation of barcode and UMI sequences from scRNA-seq reads, along with sets of equivalence classes of transcripts obtained by pseudoalignment of the reads to a reference transcriptome (Nicolae *et al.*, 2011). BUS files can be rapidly and efficiently produced, which we demonstrate by example via a novel feature in the kallisto program (http://pachterlab.github.io/kallisto/singlecell.html) that can create BUS files using data from any of seven different scRNA-seq technologies. However, the BUS format is neither technology nor software dependent. The utility of BUS files lies in their compact representation of the key information from scRNA-

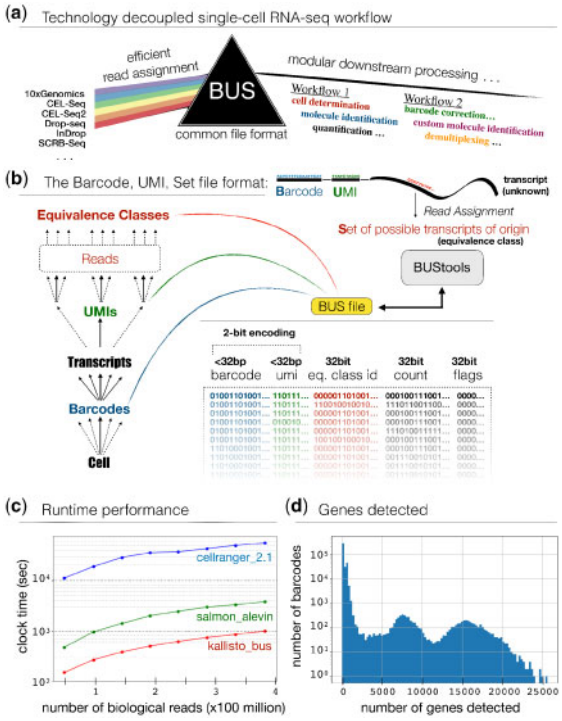

Fig. 1. (**a**) Overview of the BUS format and its applications, (**b**) description of the BUS format, (**c**) comparison of runtimes, (**d**) histogram of the number of genes detected using BUStools

seq experiments that is needed for quantification. They enable a modular approach to scRNA-seq processing that separates compute intensive read assignment from algorithmically demanding cell determination and molecule identification. Importantly, BUS also decouples technology dependencies (Fig. 1a). By virtue of avoiding the explicit representation of transcriptome sequences, BUS is also useful for sharing data in a way that removes identifying genotypes.

2 Materials and methods

The BUS format consists of two files: the first describes the mapping of equivalence classes to sets to transcripts, the second records the BUS tuples in binary format. Each record of the BUS file consists of a barcode and UMI sequence encoded using a 2-bit format, the equivalence class, count and optional flags (Fig. 1b, Supplementary Material). There is an inherent limit on the size of the barcode and UMI sequences set at 32 bp each. Each sequenced fragment corresponds to a single BUS record and no read names are stored. A BUS file can be produced by any alignment or pseudoalignment method. As a proof of principle and to illustrate the versatility and utility of BUS, we implemented a 'bus command' for the kallisto pseudoalignment software (Bray *et al.*, 2016) that will output BUS format from scRNA-seq data. 'kallisto bus' accepts as input scRNA-seq generated from 10× Genomics v1, v2 or v3 chemistry (https://support.10xgenomics.com/single-cell-gene-expression/), inDrops (Klein *et al.*, 2015), Drop-seq (Macosko *et al.*, 2015), CEL-seq/CEL-seq2 (Hashimshony *et al.*, 2016), SCRB-seq (Soumillon *et al.*, 2014) and SureCell (http://www.bio-rad.com/en-us/product/ddseq-single-cell-isolator); other technology formats can be readily processed by setting options. BUStools are a software suite developed to manipulate and organize BUS files (http://github.com/BUStools). The tools currently consist of programs for sorting BUS filesand converting BUS files to a textual representation.

3 Results

To demonstrate the utility of BUS we processed 381 992 071 scRNA-seq reads from a 1:1 mixture of fresh frozen human cells (HEK293T) and mouse cells (NIH3T3) produced with 10× Genomics technology and hosted on the 10× Genomics website (https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/hgmm_6k). 'kallisto bus' is more than 50 times faster than CellRanger, processing the dataset in 984 s versus 55 745 s with CellRanger and almost 4 times faster than the 3786 s required with Alevin (Srivastava *et al.*, 2018, Fig. 1c) using 8 cores on an Intel Xeon 6152 2.1GHz processor. Crucially, the memory requirements of 'kallisto bus' are constant in the number of reads, a feature that reduces cost for processing scRNA-seq data with cloud infrastructure. Furthermore, 'kallisto bus' is sufficiently fast with only 4 threads to provide users with limited compute resources the ability to rapidly process standard datasets in real time. To illustrate the possibilities created by the modular BUS format, we developed a simple 'getting started' post-processing notebook (available online at https://github.com/BUStools/MNP_2019). The notebook provides statistics on barcodes, UMIs and gene counts, and can be used for a rapid assessment of data. Figure 1d shows an example figure from the notebook: the distribution of genes detected for the 10× human-mouse 6k dataset.

Acknowledgements

We thank Fan Gao for helping with the benchmarking of 'kallisto bus' and BUStools. Valentine Svensson provided valuable suggestions, and we relied on his compilation of scRNA-seq read encodings (Svensson *et al.*, 2017). Jase Gehring, Lynn Yi and Tina Wang provided valuable feedback on an initial kallisto-based scRNA-seq workflow, which motivated the development of the BUS format.

Conflict of Interest: none declared.

References

Bray,N.L. *et al.* (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, **34**, 525–527.

Chen,X. *et al.* (2018) From tissues to cell types and back: single-cell gene expression analysis of tissue architecture. *Annu. Rev. Biomed. Data Sci.*, **1**, 29–51.

DePasquale,E.A.K. *et al.* (2018) CellHarmony: Cell-level matching and comparison of single-cell transcriptomes. bioRxiv, 364810.

Hashimshony,T. *et al.* (2016) CEL-Seq2: sensitive highly-multiplexed single--cell RNA-Seq. *Genome Biol.*, **17**, 77.

Klein,A.M. *et al.* (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, **161**, 1187–1201.

Macosko,E.Z. *et al.* (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, **161**, 1202–1214.

Nicolae,M. *et al.* (2011) Estimation of alternative splicing isoform frequencies from RNA-Seq data. *Algorithms Mol. Biol.*, **6**, 9.

Soumillon,M. *et al.* (2014) Characterization of directed differentiation by high-throughput single-cell RNA-Seq. bioRxiv, 003236.

Srivastava,A. *et al.* (2018) Alevin: an integrated method for dscRNA-seq quantification. bioRxiv, 335000.

Svensson,V. *et al.* (2017) Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods*, **14**, 381–387.

Svensson,V. *et al.* (2018) Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.*, **13**, 599–604.