The BUS format specification

The BUS format is a binary format for storing intermediate results for single cell RNA-Seq datasets. This repository details the specification of the format.

The motivation and example usage of the BUS format and BUStools are described in

P Melsted, V Ntranos, L Pachter, The Barcode, UMI, Set format and BUStools, bioRxiv 2018 pp: 472571.

Tools

BUS generation

- kallisto version 0.45.0 and later

BUS file manipulation

- bustools

BUS parsing and processing

- BUS R notebooks and python notebooks

Format specification

A BUS file is a binary file consisting of a header followed by zero or more BUS records. Each BUS header consists of the following elements in order

Field name	Description	Type	Value
`magic`	`fixed magic string`	`char[4]`	`BUS\0`
`version`	`BUS format version`	`uint32_t`	
`bc_len`	`Barcode length [1-32]`	`uint32_t`	
`umi_len`	`UMI length [1-32]`	`uint32_t`	
`tlen`	`Length of plain text header`	`uint32_t`	
`text`	`Plain text header`	`char[tlen]`	

BUS records are stored directly after the header in the following format, the size of each BUS record is rounded up to 32 bytes.

Field name	Description	Type
`barcode`	`2-bit encoded barcode`	`uint64_t`
`umi`	`2-bit encoded UMI`	`uint64_t`
`ec`	`equivalence class`	`int32_t`
`count`	`fragment count`	`uint32_t`
`flags`	`flags`	`uint32_t`