Using TRUmiCount

Florian G. Pflug
<florian.pflug@univie.ac.at>

Contents

1 Introduction to TRUmiCount

XXX Write Me

2 Installing TRUmiCount

2.1 Installation via Conda (Recommended)

Installing Conda

Conda is a package manager that allows easy installation of a large range of software packages. See `https://conda.io/docs/user-guide/install/index.html` for your options of how to insteall conda. Briefly, on 64-bit linux do[1]

```
INSTALLER=Miniconda2-latest-Linux-x86_64.sh
CONDA_DIR=/conda
curl -O https://repo.continuum.io/miniconda/$INSTALLER
bash $INSTALLER -p $CONDA_DIR
```

Creating an environment

Conda allows the creation of multiple *environments*, each containing different collections of packages. We will now create an environment for TRUmiCount

```
$CONDA_DIR/bin/conda create -n trc
```

This environment is now *activated* to make it the target of further conda commands, and the installed software visible. This must be done every time a new terminal window is opened!

```
source $CONDA_DIR/bin/activate trc
```

Installing BioConda

Conda packages are organized into so-called *channels*. We add the BioConda channel which provides many common tools for dealing with high-throughput sequencing data

```
conda config --env --add channels defaults
conda config --env --add channels conda-forge
conda config --env --add channels bioconda
```

Installing TRUmiCount

Finall we add the channel that supplies TRUmiCount and a modified version of umi_tools[2] with improved handling of paired-end reads[3]

```
conda config --env --add channels \
  http://tuc:tuc@www.cibiv.at/~pflug_/trumicount/
```

TRUmiCount and our version of umi_tools can now be installed

```
conda install TRUmiCount umi_tools samtools
```

[1] Instead of `/conda`, you can choose any other directory to install conda into

[2] `https://github.com/CGATOxford/UMI-tools`

[3] Note that the backslash (“\”) only serves to make your shell ignore the linebreak that follows it. If you enter the command as a single line, skip the backslash

2.2 Manual Installation

XXX Write Me

3 Using TRUmiCount

3.1 Supported input formats & options

To be able to separate true UMIs from biases and to estimate and correct for the percentage of true UMIs that are lost during library preparation or data processing, TRUmiCount analyses the distribution of read counts per UMI for each gene (or any other type of genomic feature). TRUmiCount by default assumes that

- Each UMI initially had two copies. This is e.g. the case of molecules before amplification were double-stranded and the copies produced from both strands are identical. This number can be changed with "`--molecules COPIES`"
- UMIs must be supported by at least two reads to be assumed to be a true UMI and not a phantom. This threshold can be changed with "`--threshold TH`".

Reading UMIs from a BAM File

If a mapped BAM File is provided as input with "`--input-bam BAMFILE`", TRUmiCount uses umi_tools's `group` tool to extract a list of UMIs and their read counts from a BAM file. Sequencing errors in the UMIs are corrected by umi_tools by *merging* similar UMIs into one. When reading BAM file, on top of the defaults mentioned above, TRUmiCount assumes that

- The BAM file must have a corresponding index. A suitable index can be created with "`samtools index BAMFILE`".
- The sequence name corresponds to the gene name. Alternatively, the gene names can be stored in BAM file tags – this can be changed by using umi_tool's `--gene-tag=GENE_TAG` option. To tell TRUmiCount to invoke umi_tools using that option, use "`--umitools-option --gene-tag=GENE_TAG`" when invoking TRUmiCount.
- The UMI was appended to the read name, and separated by ":". A different separator can be specified with "`--umi-sep SEPARATOR`"
- The BAM file contains single-end reads (read2 is ignored). To take the mapping position of both mates into account when grouping reads by UMI, specify "`--paired`".
- Reads with a mapping quality below 20 should be ignored. This threshold can be changed with "`--mapping-quality MAPQ`".

Reading UMIs from a tab-separated file

Instead of using umi_tools to extract UMI and their read counts from a BAM file, TRUmiCount can read a previously computed table of UMIs with "`--input-umis`". The table must be tab-separated with one row per UMI and contain at least the columns "sample", "gene", "reads". When dealing with strand UMIs (see section *Strand UMIs*), TRUmiCount will also use columns "pos" and "end" containing the mapping position of read1 respectively read2.

The umi_tools and BAM-related options "`--umi-sep`", "`--umitools`", "`--umitools-option`", "`--paired`" and "`--mapping-quality`" are ignored if "`--input-umis`" is used.

The option "`--output-umis`" (together with "`--input-bam`") produces a suitable input file for "`--input-umis`". This can be used to avoid the overhead of running umi_tools multiple times if the same input BAM file is processed multiple times with TRUmiCount, e.g. to test different read count thresholds or initial molecule counts.

3.2 Output

XXX Write Me

3.3 Strand UMIs

Some UMI-based library preparation protocols produce *strand UMIs* where the two strands of an initial double-stranded template molecule produce distinct (but related) UMIs. Filtering out UMIs for which the partner UMI corresponding to the second strand is not detected offers second possibility (besides the read count threshold) for filtering our phantom UMIs.

TRUmiCount supports stranded UMIs as produced by the protocol of Shiroguchi *et al.*[4]. With this protocol, both read1 and read2 carry a separate molecular barcode. UMI pairs belonging to the same double-stranded template molecule are found by looking for pairs of UMIs whose read1 and read2 barcodes and mapping positions are swapped (mapping position here refers to the genomic coordinate of the first mapped base in *read direction*, i.e. for reverse-mapped reads this differes from the mapping position stated in the BAM file).

When working with strand UMIs, the initial molecule count should usually be set to 1 (the default is 2!), i.e. "`--molecules 1`" should be used.

Filtering out incomplete strand UMI pairs

With the option "`--filter-strand-umis`", UMIs are filtered out if their partner UMI cannot be detected. Note that the actual loss rate in this mode is not simply the probability $\mathbb{P}(C < T)$ of an UMI having fewer than T reads.

[4] Shiroguchi, K., Jia, T. Z., Sims, P. A. & Xie, X. S. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 1347-1352 (2012).

Since we filter out UMIs *either* if they themselves have fewer than T reads, *or* if their partner UMI wasn't detected at all, the actual loss is $1 - (1 - \mathbb{P}(C < T)) \cdot (1 - \mathbb{P}(C = 0))$. TRUmiCount adjust the loss computation accordingly and outputs corrected losses. The model distribution shown in the plots, however, does not (and cannot) take this adjustment into account, so that the stated loss is no longer simply the sum of the model probabilities for read counts less than T.

Combining strand UMIs into pairs

With the option "`--combine-strand-umis`", partner UMIs are paired, and the read count threshold is applied to be partners. In this mode, the actual loss (i.e. fraction of true UMIs removed by the filter) is $1 - (1 - \mathbb{P}(C < T))^2$, because we now filter out UMI pairs if *either* of the partners have fewer than T reads. Againm TRUmiCount adjust the loss computation accordingly and outputs corrected losses. And again the model distribution shown in the plots, does not (and cannot) take this adjustment into account, so that the stated loss is no longer simply the sum of the model probabilities for read counts less than T.

4 Parameter Reference

XXX Write Me

5 Examples

The example data can be downloaded from `http://tuc:tuc@www.cibiv.at/~pflug_/trumicount`. Before TRUmiCount can process these BAM files, they need to be indexed by running

```
samtools index sg_100g.bam
samtools index kv_1000g.bam
```

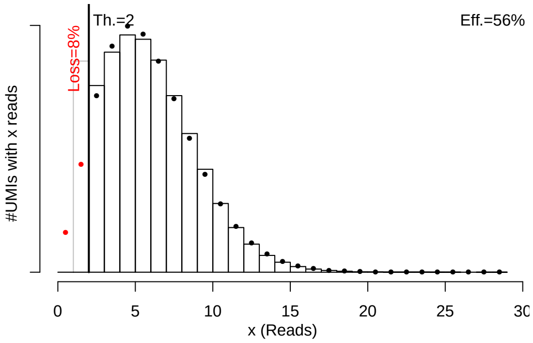
5.1 Single-End Data

The file `kv_1000g.bam` contains a reduced (restricted to the first 0100 genes, and subsamples to 50%) version of data published by Kivioja *et al.*[5].

```
trumicount --input-bam kv_1000g.bam \
  --molecules 2 --threshold 2 \
  --output-plot kv_1000g.pdf --plot-x-bin 1 \
  --output-counts kv_1000g.tab \
  --cores 4
```

[5] Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods* **9**, 7274 (2011)

This produces the following diagnostic plot kv_1000g.pdf showing the observed distribution of reads per UMI and our model's predicted distribution and loss.



5.2 Paired-End Data with strand UMIs

The file sg_100g.bam contains a reduced (restricted to the first 100 genes, and subsamples to 25%) version of data published by Shiroguchi *et al.*[6] which uses *strand UMIs*.

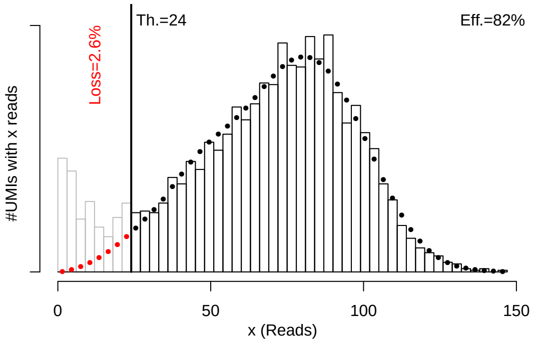
Filtering out incomplete strand UMI pairs

The data from from Shiroguchi *et al.* is analyzed in paired-end mode, and UMIs are filtered-out if the UMI corresponding to the second strand of the original template molecule is not detected.

```
trumicount --input-bam sg_100g.bam --umipair-sep '-'\
  --paired --filter-strand-umis --molecules 1 --threshold 24 \
  --output-plot sg_100g.pdf --plot-x-bin 3 \
  --output-counts sg_100g.tab
```

This produces the following diagnostic plot sg_100g.pdf showing the observed distribution of reads per UMI and our model's predicted distribution and loss. Note that for the reasons explained in *Strand UMIs* the stated loss deviates from the sum of the model probabilities for read counts less than T.

[6] Shiroguchi, K., Jia, T. Z., Sims, P. A. & Xie, X. S. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 1347-1352 (2012).

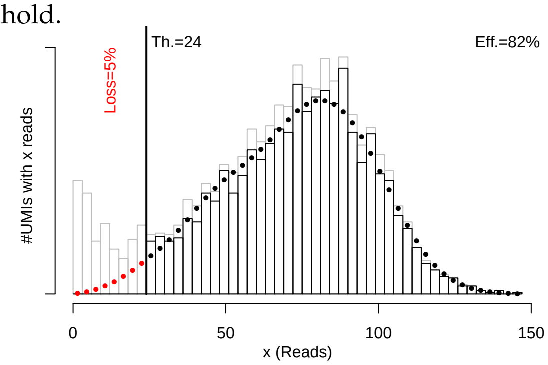


Combining strand UMIs into pairs

The data from from Shiroguchi *et al.* is analyzed in paired-end mode, this time combining UMIs stemming from the two strands of a single template molecule and applying the read count threshold to both partner's read counts. UMIs without a mate to combine with are dropped (as they are for "`--filter-strand-umis`").

```
trumicount --input-bam sg_100g.bam --umipair-sep '-'\
  --paired --combine-strand-umis --molecules 1 --threshold 24 \
  --output-plot sg_100g_comb.pdf --plot-x-bin 3 \
  --output-counts sg_100g_comb.tab
```

This produces the following diagnostic plot `sg_100g_comb.pdf` showing the observed distribution of reads per strand UMI and our model's predicted distribution and loss. Note that for the reasons explained in *Strand UMIs* the stated loss deviates from the sum of the model probabilities for read counts less than T. Also note that larger number of phantom UMIs to the *right* of the read count threshold – those are UMI pairs with one read count above and one below the threshold.


Th.=24
Eff.=82%
Loss=5%
#UMIs with x reads
x (Reads)