BioIT guidelines

Bioinformatics guide for LACEseq/ALL-IN-ONE RiboLace Gel Free/ALL-IN-ONE RiboLacePro kit

FOR RESEARCH USE ONLY

Supplemental product information and tips for success

Expected Illumina sequencing output.

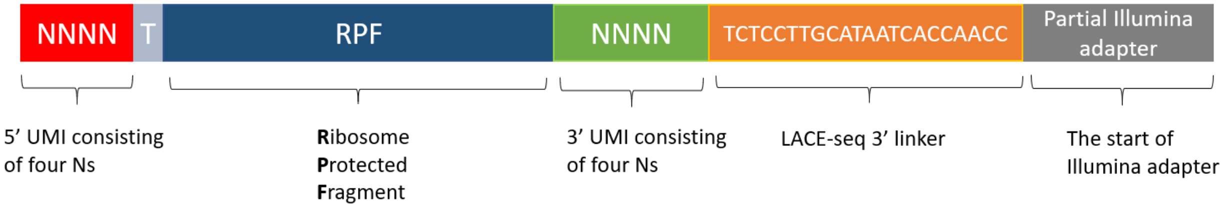
NNNN
T
RPF
NNNN
TCTCCTTGCATAATCACCAACC
Partial Illumina adapter
5' UMI consisting of four Ns
Ribosome Protected Fragment
3' UMI consisting of four Ns
LACE-seq 3' linker
The start of Illumina adapter

Figure 1. Expected Illumina sequencing output: example of a read generated.

Unique molecular identifiers (UMIs) are strings of random nucleotides that are attached to RPFs prior to PCR amplification and can be used to accurately detect PCR duplicates.

The T at position fifth precedes the start of the RPF. The sequence content of a high-quality library has a T peak in position fifth in 90-100% of the reads (Figure 2).

To check T peak use fastqc command:

- fastqc --outdir outputdir input.fastq

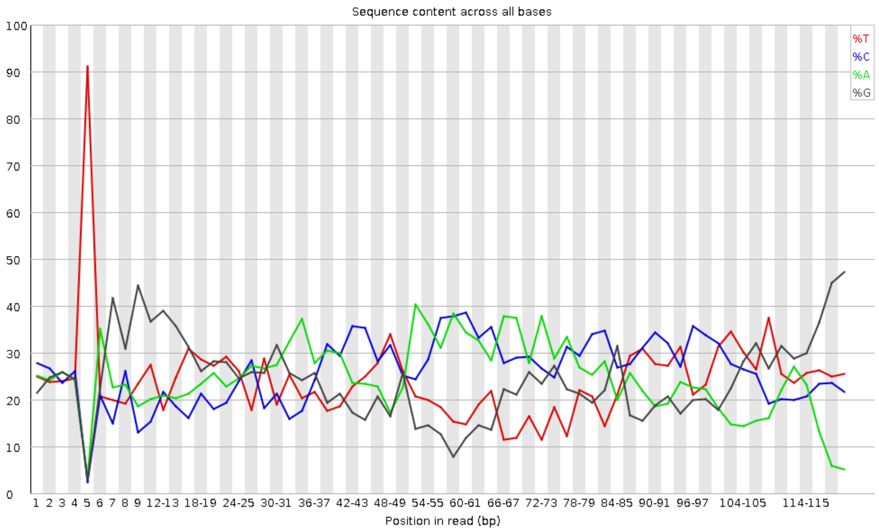

Figure 2. Sequence content across all bases graph. The sequence content of a high-quality library has a T peak in position five in 90-100% of the reads.

Workflow overview

There are 5 main steps in the analysis pipeline:

E1. Software installation
E2. Trimming/UMI extraction
E3. Filtering rRNA, tRNA and ncRNA
E4. STAR alignment

E5. RiboWaltz pipeline

Step E1. Software installation:

Information and guides to install the required tools. Though more recent versions of the programs will also be compatible with this pipeline, the workflow is intended to function with the versions listed:

- Dependencies
  - Trimming:
    - Cutadapt (https://cutadapt.readthedocs.io/en/stable/installation.html)
  - Quality Control:
    - Fastqc (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)
  - Alignment:
    - bowtie2 (https://bowtie-bio.sourceforge.net/bowtie2/index.shtml)
    - STAR (https://github.com/alexdobin/STAR)
  - Utilities:
    - umi_tools (https://github.com/CGATOxford/UMI-tools)
    - samtools (https://www.htslib.org/)
  - Ribosomal Footprint Analysis:
    - RiboWaltz (R) (https://github.com/LabTranslationalArchitectomics/riboWaltz)
- Build Aligner Indexes
  - To build bowtie2 indexes fasta files of tRNAs, rRNAs and snRNAs or ncRNAs are necessary. You can find those files https://rnacentral.org/.
  - To build STAR index also gtf file is needed. And those files can be found at https://www.gencodegenes.org/ and https://www.ensembl.org

Once the tools have been installed, you will need to make sure that the UNIX environment variables are appropriately set. You can either add the location of the executables installed to your PATH variable or create a new directory called bin in your home directory, copy the executables to this location, and add the location of the bin directory to your PATH variable.

To change your PATH variable, enter (assuming bash shell):

```
    > export PATH = <list of paths>:$PATH
```

Parameter	Definition
PATH = <list of paths>:$PATH	specify number of threads in computer for this job (Depends on the computer)

Step E2: Trimming/UMI extraction

Proper trimming of the reads is important for efficient mapping. Here we provide some guidance on the use of (E2.1) cutadapt (Martin M. 2011) to remove Linker MC+ (MC+), (E2.2) UMI-tools extract (Smith T. 2017) to move the UMI sequence from the read to the read name so that PCR duplicates can be removed after the alignment, (E2.3) cutadapt to remove the T preceding the RPF.

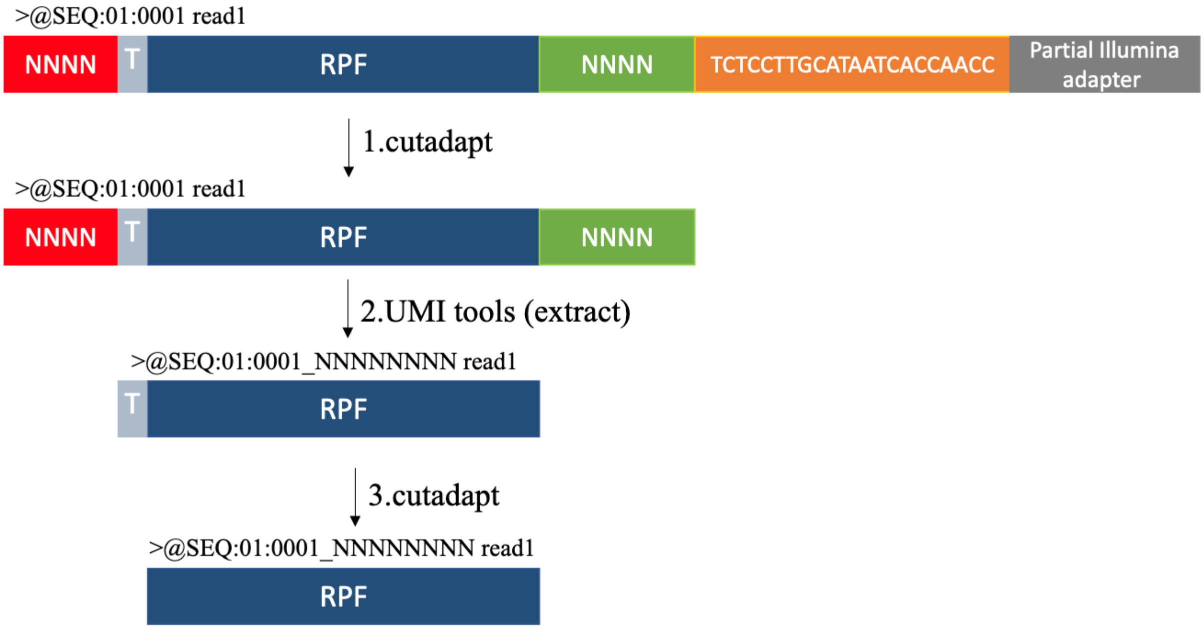

Figure 3. UMI extraction and trimming step schematic. The RPF extraction is done in 3 steps: linker removal, UMI extraction, T removal.

E2.1: cutadapt

First the Linker MC+ (MC+) is trimmed from the 3’ end of each read and only reads longer than X+9 nt are retained, while shorter reads are discarded:

```
cutadapt --cores N --minimum-length X+9 -a TCTCCTTGCATAATCACCAACC --
discard-untrimmed -o trim.fastq input.fastq
```

Parameter	Definition
`--cores N`	specify number of threads in computer for this job (Depends on the computer)
`--minimum-length X+9`	Reads are retained if they are longer than X+9 nt, where X is the length of the RPF (usually X=20 for ribosome profiling analysis), and 9 is the sum of the lengths of the 5’ and 3’ UMIs
`-a TCTCCTTGCATAATCACCAACC`	Removal of the LACE-seq 3’ linker and any sequence that may follow
`--discard-untrimmed`	Reads in which *no* adapter is found are discarded
`-o trim.fastq`	The output file name
`input.fastq`	The input file name

E2.2: UMI-tools (extract)

The sequence of the 5’ and 3’ UMIs are moved from the read sequence to the read name:

```
umi_tools extract -I trim.fastq --bc-
pattern='^(?P<umi_1>.{4}).+(?P<umi_2>.{4})$' --extract-method=regex -S
extract.fastq --log=<umi_extract.log>
```

Parameter	Definition
`-I trim.fastq`	The input file name must be the same as the output file name in step1
`--bc-pattern='^(?P<umi_1>.{4}).+(?P<umi_2>.{4})$'`	extract the first 4 (5’UMI) and the last 4 bases (3’UMI) of each read
`--extract-method`	defines method for UMI extraction
`-S extract.fastq`	The output file name

NOTE: UMI-tools dedup can be used after alignment to remove duplicates based on the mapping coordinate and the UMI attached to the read name.

E2.3: cutadapt

The T preceding the RPF is then removed:

```
cutadapt --cores N -g ^T --discard-untrimmed -o trim2.fastq
extract.fastq
```

Parameter	Definition
`--cores N`	specify number of threads in computer for this job (Depends on the computer)
`-g ^T`	Removal of the first T at the start of each read
`-o trim2.fastq`	The output file name
`extract.fastq`	The input file name must be the same as the output file name in step E2.2

Step E3: Filtering rRNA, tRNA and ncRNA

In order to remove and quantify ribosomal RNA (rRNA) content or other contaminants (tRNAs and snRNAs etc) in your sample prior to alignment to the genome, you can align the trimmed reads against specific contaminant sequences. The first step in removing contaminants is to create a FASTA formatted file containing contaminating sequences from your sample to align against, using the Bowtie aligner (Bowtie2-build https://bowtie-bio.sourceforge.net/bowtie2/index.shtml).
To build bowtie2 indexes fasta files of tRNAs, rRNAs and snRNAs or ncRNAs are necessary. You can find those files https://rnacentral.org/.

```
bowtie2-build --threads N - f <reference.fasta.file>
<given_index_name>
```

Parameter	Definition
`--threads N`	specify number of threads in computer for this job (Depends on the computer)
`- f <reference.fasta.file> <given_index_name>`	f: specify fasta file location and name (Eg: /go/to/reference.fa) and given_index_name refers to the location and name of the indexes (Eg: /go/to/index/rRNA)

E3.1: removing rRNA contaminant

```
 bowtie2 --threads N -N 1 --no-1mm-upfront -q <trimmed.fastq.gz> --
un=<norRNA.fastq.gz> -x <rRNA_bowtie_index>
```

Parameter	Definition
`--threads N`	specify number of threads in computer for this job (Depends on the computer)
`-N 1`	Number of allowed mismatches
`--no-1mm-upfront`	This option prevents Bowtie 2 from searching for 1-mismatch end-to-end alignments
`-q <trimmed.fastq.gz>`	Input filename
`--un=<norRNA.fastq.gz>`	output not aligned reads
`-x <rRNA_bowtie_index>`	Index file for alignment

E3.2: removing tRNA contaminant

```
bowtie2 --threads N -N 1 --no-1mm-upfront -q <norRNA.fastq.gz> --
un=<norRNA_notRNA.fastq.gz> -x <tRNA_bowtie_index>
```

Parameter	Definition
`--threads N`	specify number of threads in computer for this job (Depends on the computer)
`-N 1`	Number of allowed mismatches
`--no-1mm-upfront`	This option prevents Bowtie 2 from searching for 1-mismatch end-to-end alignments
`-q <trimmed.fastq.gz>`	Input filename

`--un=<norRNA_notRNA.fastq.gz>`	output not aligned reads
`-x <tRNA_bowtie_index>`	Index file for alignment

E3.3: removing ncRNA contaminant

```
 bowtie2 --threads N -N 1 --no-1mm-upfront -q
<norRNA_notRNA.fastq.gz> --un=<norRNA_notRNA_noncRNA.fastq.gz> -x
<ncRNA_bowtie_index>
```

Parameter	Definition
`--threads N`	specify number of threads in computer for this job (Depends on the computer)
`-N 1`	Number of allowed mismatches
`--no-1mm-upfront`	This option prevents Bowtie 2 from searching for 1-mismatch end-to-end alignments
`-q <trimmed.fastq.gz>`	Input filename
`--un=<norRNA_notRNA_noncRNA.fastq.gz>`	output not aligned reads
`-x <ncRNA_bowtie_index>`	Index file for alignment

Step E4: STAR Alignment

The next step for analysis is to align the remaining reads to the genome using the STAR (https://github.com/alexdobin/STAR).

To build STAR index also gtf file is needed. And those files can be found at https://www.gencodegenes.org/ and https://www.ensembl.org

```
STAR --runMode genomeGenerate --runThreadN N --genomeDir
<location_for_index> --genomeFastaFiles <location_of_fasta_file>
--genomeSAindexNbases <calculated size> --sjdbGTFfile
<location_of_gtf_file>
```

Parameter	Definition
`--runMode genomeGenerate`	option directs STAR to run genome indices generation job
`--runThreadN N`	specify number of threads in computer for this job (Depends on the computer)
`--genomeDir <location_for_index>`	location_for_index: refers to the location and name of the indexes
`--genomeFastaFiles <location_of_fasta_file>`	location_of_fasta_file: specifies one or more FASTA files with the genome reference sequences. The tabs are not allowed in chromosomes' names, and spaces are not recommended.
`--genomeSAindexNbases <calculated size>`	genomeSAindexNbases: can be find with; min (14, log2(GenomeLength)/2-1) for hg38 genome its min (14, log2(3272116950)/2-1) = 14
`--sjdbGTFfile <location_of_gtf_file>`	location_of_gtf_file: specifies the path to the file with annotated transcripts in the standard GTF format.

Step E5: RiboWaltz pipeline

For the RiboSeq Quality Metrics analysis you can use RiboWaltz, an R package that integrates quality controls of the ribosome profiling data, P-site identification for improved interpretation of positional information and a variety of graphical representations.

Use transcriptome BAM file and GTF annotation file to run riboWaltz (https://github.com/LabTranslationalArchitectomics/riboWaltz).

Contacts

Info

info@immaginabiotech.com

Sale support (quoting, ordering, and order status update)

orders@immaginabiotech.com

Technical service (technical inquiries and quality complaints)

techsupport@immaginabiotech.com

Viale Dell’industria, 47, 38057, Pergine Valsugana (TN), ITALY

https://www.immaginabiotech.com/

+39 04611787270

Notes: