ANALYSIS OF TCP-seq in RIBOGALAXY

In this tutorial we are going to show how TCP-seq data can be processed in the RiboGalaxy instance available at https://ribogalaxy.genomicsdatascience.ie/.

TCP-seq - Translation Complex Profile Sequencing originally developed in yeast [PMID: 27437580]. This type of sequencing can capture footprints that are protected by 40S and 80S ribosomal subunits. Its modification - selective footprinting (or Sel-TCP-seq) includes an immunoprecipitation step to assay the location of both 40S and 80S ribosome complexes containing proteins of interest (e.g. translation initiation factors). These methods allow us to study translation dynamics and decipher which protein factors play an important role in translation steps.

Here we assume that you already used RiboGalaxy for preprocessing raw standard Ribo-seq data and mapping to genome and transcriptome (part 1,2,3 of the tutorial) and therefore we only provide screenshots for some of the steps and the final result.

Input TCP-seq Ribo-Seq Reads

We are going to use a subsample of human 40S footprint reads from [PMID: 32589966] study, SRA id: SRR10346283. This subsample contains only selected reads that map to a set of transcripts including ENST00000559916 (B2M), ENST00000371222 (JUN), ENST00000621592 (MYC), ENST00000373316 (PGK1), ENST00000674681 (ACTB), ENST00000396861 (GAPDH). It's called '***TCP_human.fq***'. It is available in the same archive where tutorials are stored or alternatively you can download it from here: https://www.dropbox.com/s/j2ett6wwjmf7j87/TCP_human.fastq?dl=0.

They used a kit where 5'end of the read contains extra 3nt that need to be trimmed and 3'end of a read contains 'polyA' tail derived from oligo dT primer followed by an adapter. Typically, information about the protocol can be found in the corresponding paper.

Here how read looks like:

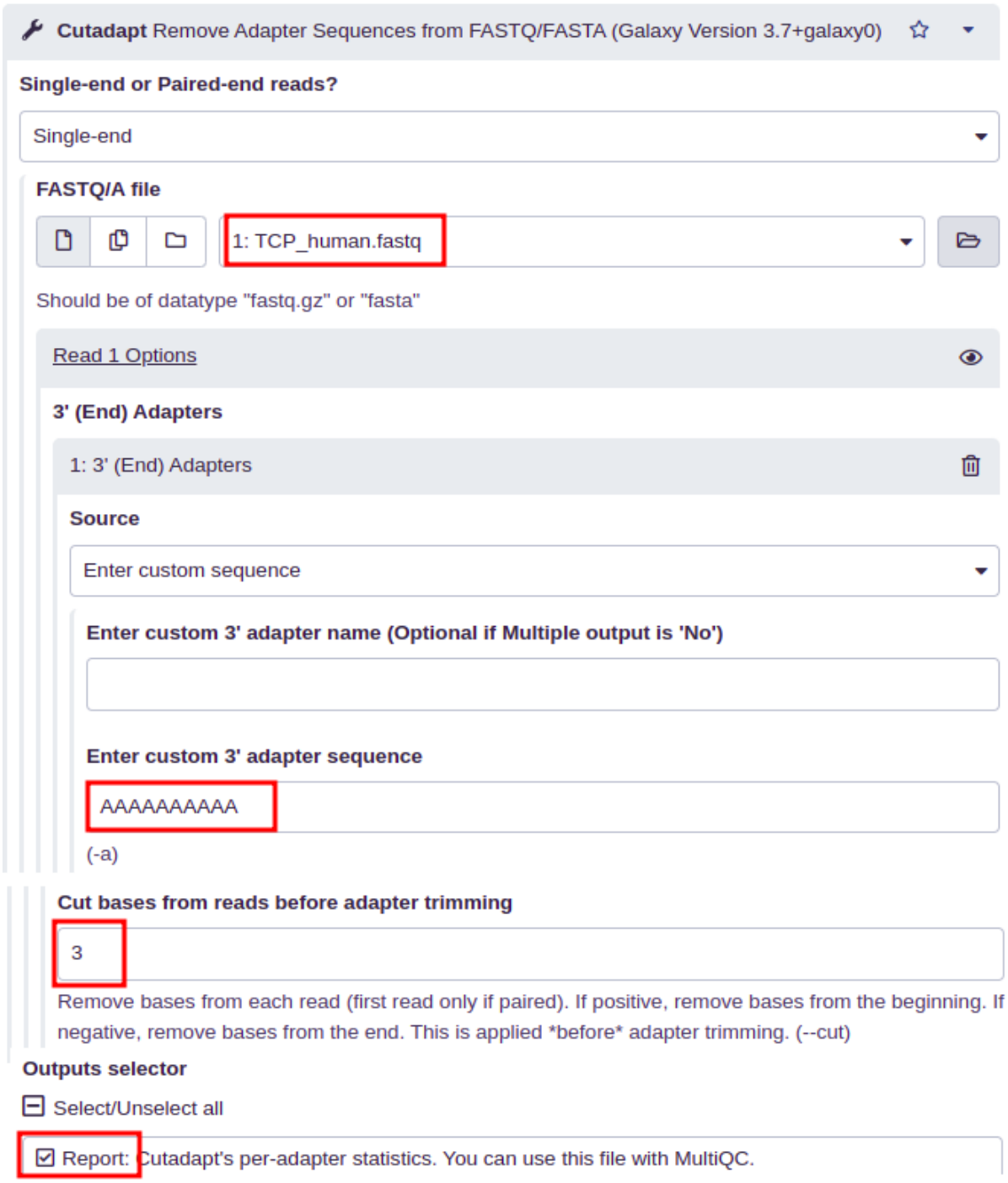
5' **UUU** - RPF - **AAAAAAAAAA** - adapter 3'

UUU - untemplated additions
RPF - ribosome protected fragment
AAA - 'polyA' adapter
adapter - Illumina adaptor

Processing steps

Here are the steps you need to do to obtain footprints from raw reads (additionally you can check the quality of the reads using **Fastqc** tool from the **Preprocessing** section):

- After raw data is uploaded (use ‘fastqsanger’ file type and hg38 as Genome), we will trim adapters and untemplated additions using ***Cutadapt***. We can also ask for a report. Choose ‘AAAAAAAAAA’ as 3’end adapter and ‘3’ nt to be trimmed from 5’end.



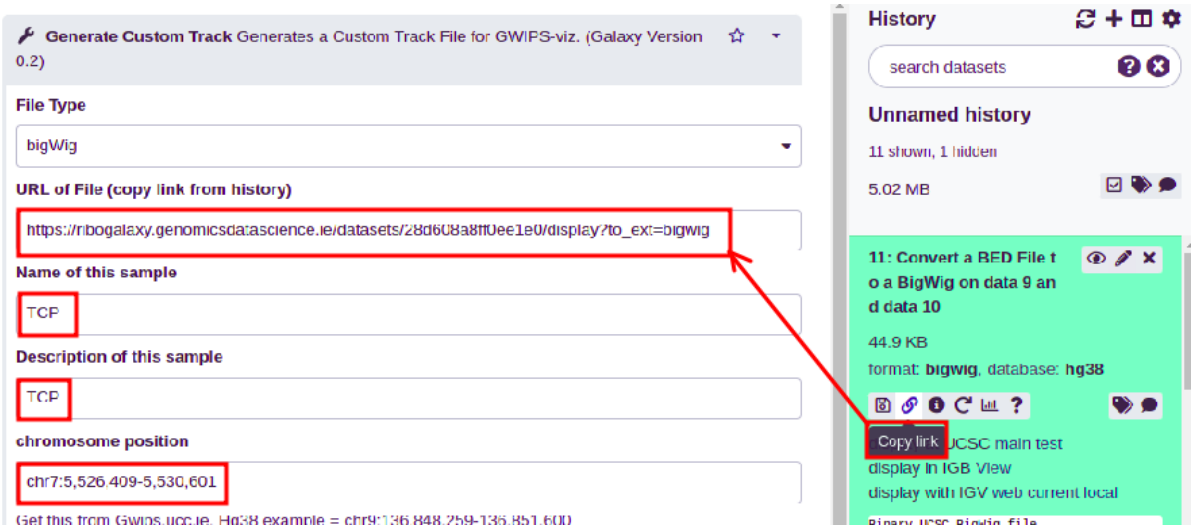
Now we can take a look at the report, almost all reads contain adapters:

```
Total reads processed:                  7,183
Reads with adapters:                    7,177 (99.9%)
```

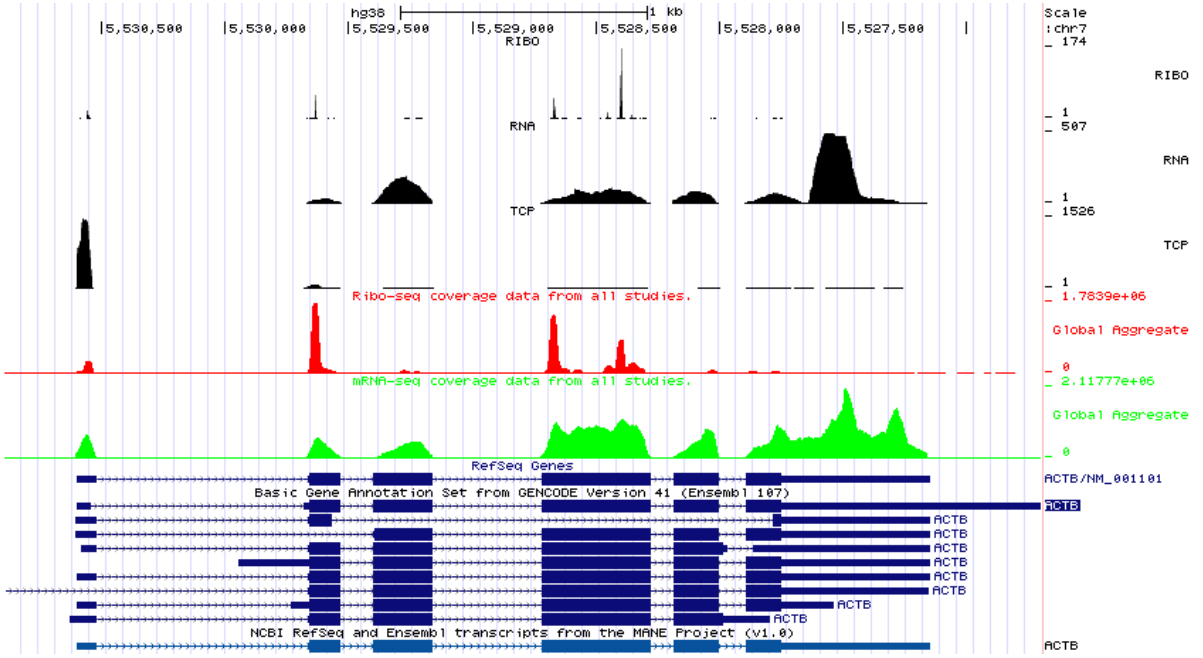
- Then we will deal with typical major contaminants of ribosome profiling - tRNA and rRNAs by using ***Bowtie ncRNA Removal***.

Steps for mapping to the genome:

- First, we will map clean reads (ribosome protected fragments) to the genome using ***Bowtie Genome Alignment***. We will use hg38 as reference.
- Then we sort alignments based on coordinates using ***Samtools sort***.
- We need to obtain chromosome sizes by using the Get ***Chromosome Sizes tool*** (also using hg38).
- Using the sorted BAM file, we will create a coverage file in BED format using **BedTools Genome Coverage.**
- Next we will convert the BED file to BigWig so that it can be uploaded and visualised in GWIPS-viz by using ***Convert a BED File to a BigWig*** tool***.***
- In order to upload the resulting BigWig track file to GWIPS-viz, we will need to create a link using the Generate **Custom Track** tool. You will need to copy a link to the BigWig file and use it as input. Add name and description of the sample (TCP), as well as any chromosome position of interest, e.g. chr7:5,526,409-5,530,601 (it can be changed in genome browser). Click 'Execute'. This tool will output a file containing a link. You need to download this file and then upload to the GWIPS-viz browser.



This is how the sample (track called TCP) looks in GWIPs-viz after processing. You may notice that most of the data is localised is 5'leader.

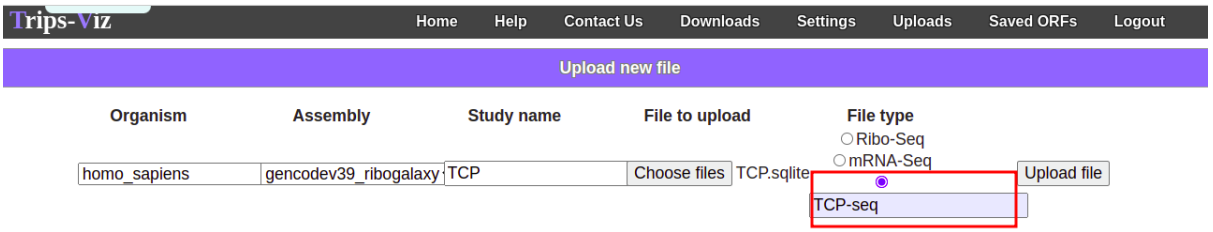


Here, instead of calculating A- or P-sites as it's done for regular Ribo-seq footprints, for small subunit (40S) footprints we calculated genome coverage (similar to RNA-seq). By genome coverage here we mean the number of read nucleotide bases aligned to a specific locus in a reference genome (https://bedtools.readthedocs.io/en/latest/content/tools/genomecov.html).

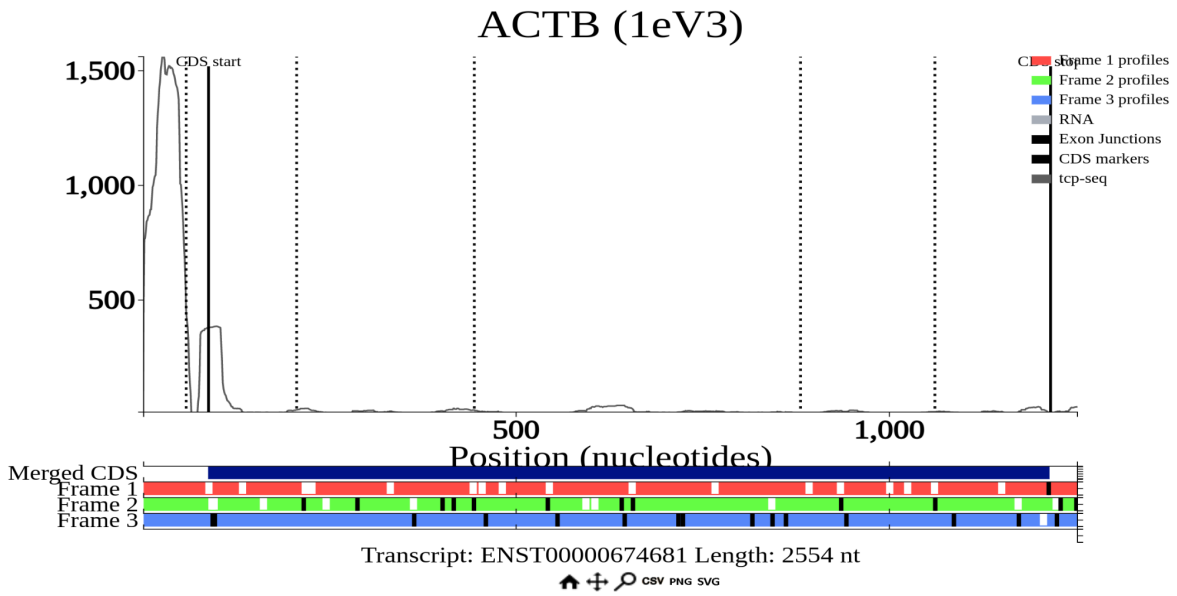
Scanning ribosome moves by 1nt steps unlike the 80S ribosome whose translocation results in 3nt (codon) steps thus leading to a triplet periodicity signal. Also, 40S footprints vary greatly in length unlike 80S translating ribosomes.

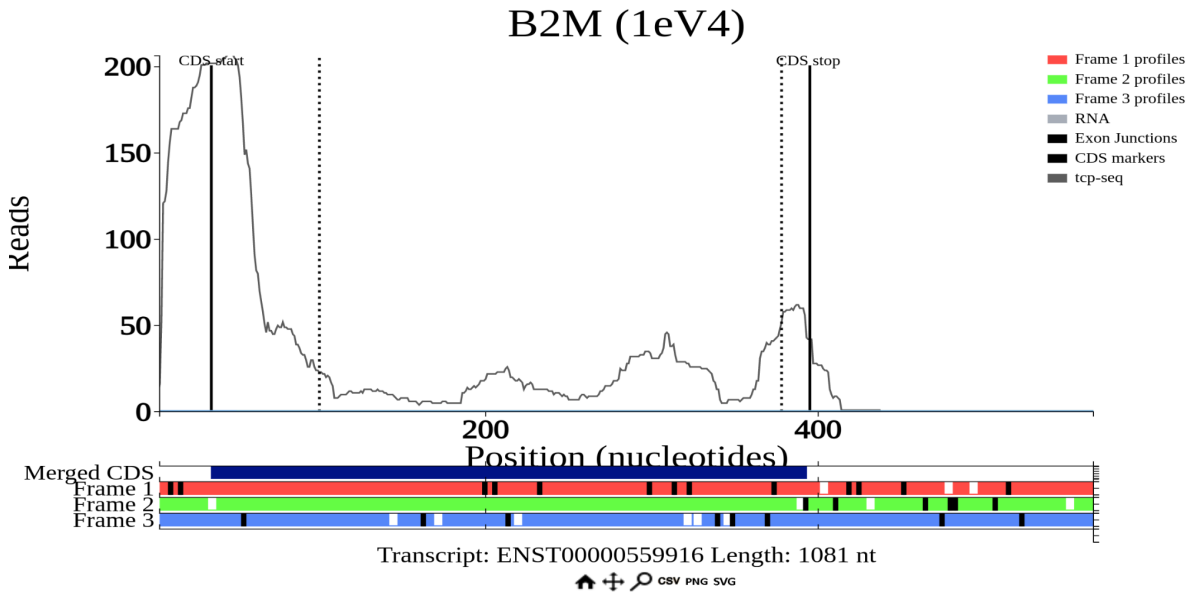
Steps for mapping to the transcriptome:

- First, we will map clean reads (footprints) to the transcriptome (GENCODE v39) using ***Bowtie Transcriptome Alignment.***
- Next, we will sort the BAM file by name using ***Samtools sort*** since this order is required for building SQLITE files. SQLITE is a format that is used to store mapped reads and associated statistics (e.g. triplet periodicity and ambiguity of mapping) for downstream analysis and visualisation in Trips-viz browser.
- Finally, we will create a SQLITE file using the ***BAM to Sqlite*** tool (GENCODE v39). We will then upload it to Trips-viz and explore TCP-seq profiles. When uploading the file to the Trips-viz, you need to select 'Other' as File Type and put 'TCP-seq':



This is how 40S-profiles looks like in Trips-viz (most of the footprints are accumulated in 5'leaders of transcripts):





ANALYSIS OF TI-seq in RIBOGALAXY

Now we are going to show the example of processing TI-seq data in RiboGalaxy instance available at https://ribogalaxy.genomicsdatascience.ie/.

Translation initiation sequencing (TI-seq) allows mapping translation initiation sites in the genome. While regular Ribo-seq uses cycloheximide treatment (a translation elongation inhibitor), TI-seq employs lactimidomycin [PMID: 22927429] or harringtonine [PMID: 22056041] treatment for predominant capture of initiating ribosomes.

Here we assume that you already used RiboGalaxy for preprocessing raw standard Ribo-seq data and mapping to genome and transcriptome (part 1,2,3 of the tutorial) and therefore we only provide screenshots for some of the steps and the final result. There is not much of a difference in processing regular Ribo-seq data and TI-seq data in RiboGalaxy.

Input TI-seq Ribo-Seq Reads

We created a subsample of TI-seq data from study [PMID: 26900662, SRR1802150]; human fibroblast cells treated with harringtonine. This subsample contains only selected reads that map to a set of transcripts including ENST00000559916 (B2M), ENST00000371222 (JUN), ENST00000621592 (MYC), ENST00000373316 (PGK1), ENST00000674681 (ACTB), ENST00000396861 (GAPDH). It's called '***INIT_human.fastq***'. It is available in the same archive where tutorials are stored or alternatively you can download it from here: https://www.dropbox.com/s/fzl2ra19o0xcgj0/INIT_human.fastq?dl=0.

Here how read looks like:

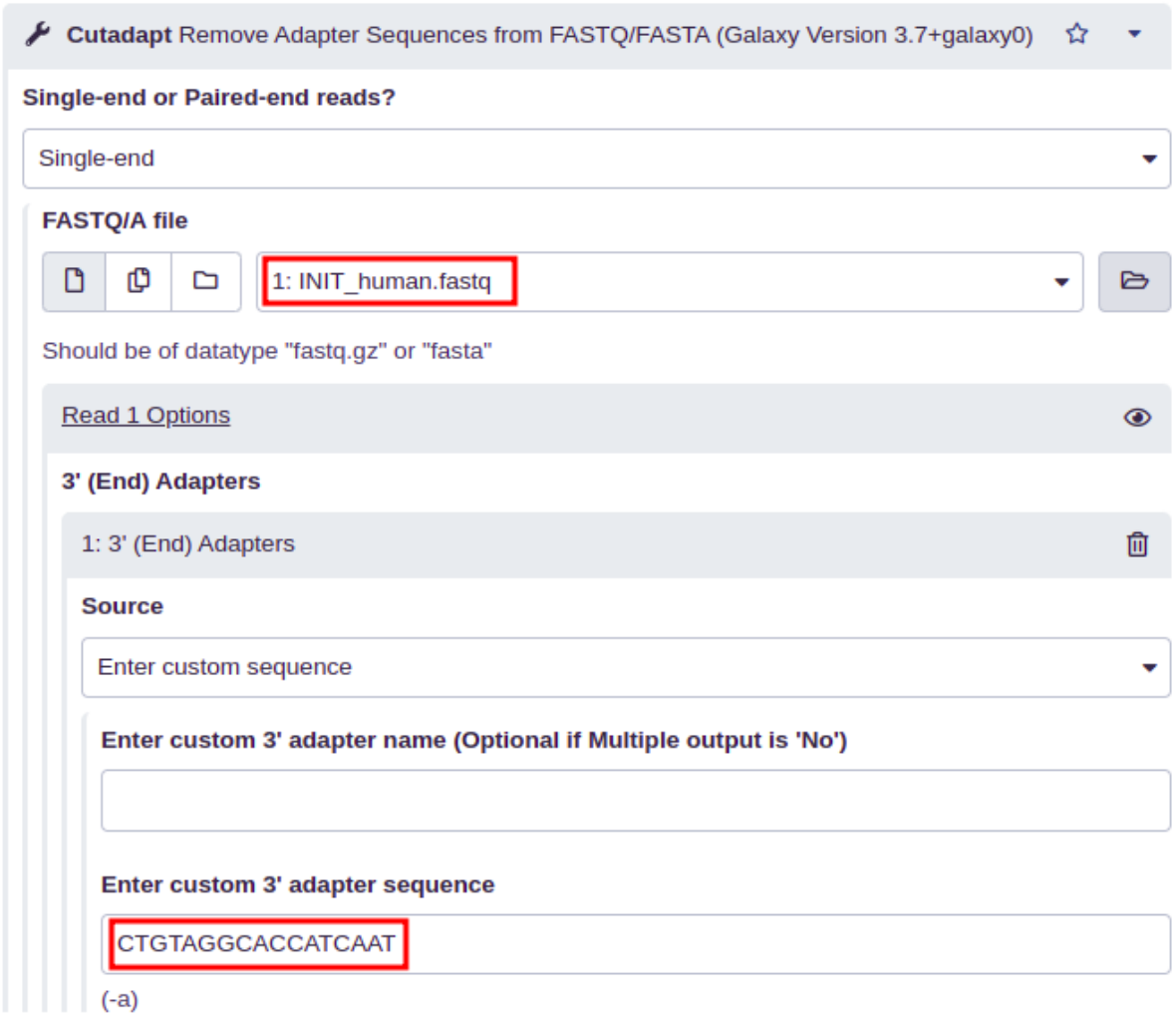
5' RPF - adapter 3'

Where:
RPF - ribosome protected fragment
adapter - CTGTAGGCACCATCAAT

Processing steps

Here are the steps you need to do to obtain footprints from raw reads (additionally you can check the quality of the reads using **Fastqc** tool from the **Preprocessing** section):

- After raw data is uploaded (use ‘fastqsanger’ file type and hg38), we will trim adapters, untemplated additions and barcodes using ***Cutadapt***. We can also ask for a report. Choose ‘CTGTAGGCACCATCAAT ’ as a 3'end adapter. We also ask for the report here.



All reads contain adapters:

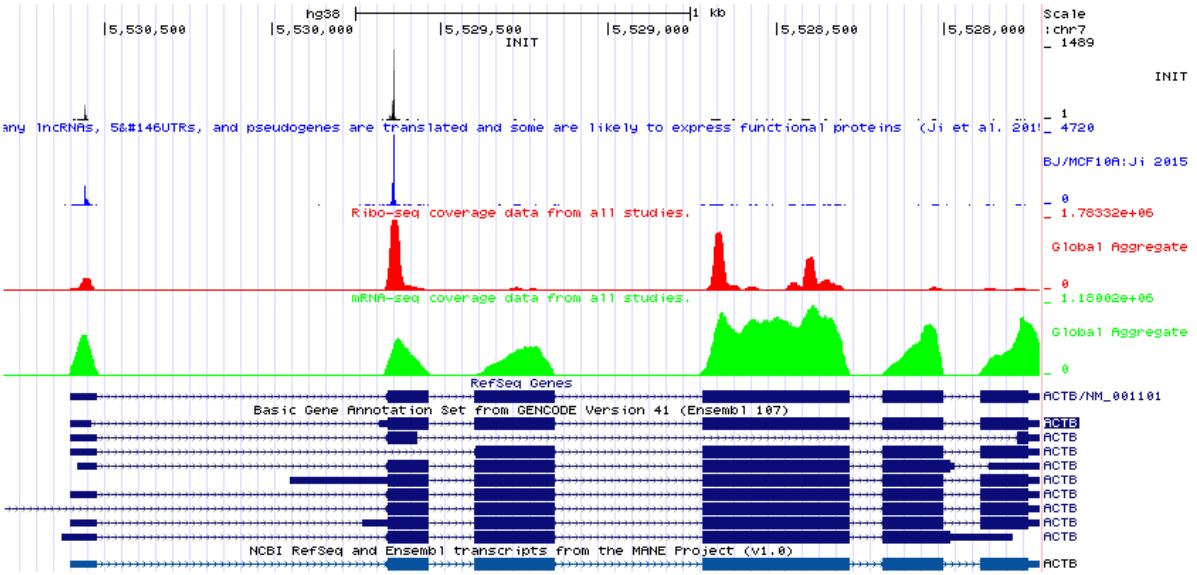
```
=== Summary ===

Total reads processed:                  17,611
Reads with adapters:                    17,611 (100.0%)
```

- Then we will deal with typical major contaminants of ribosome profiling - tRNA and rRNAs by using ***Bowtie ncRNA Removal***.

Steps for mapping to the genome:

- First, we will map clean reads (ribosome protected fragments) to the genome using ***Bowtie Genome Alignment***. We will use hg38 as reference.
- Then we sort alignments based on coordinates using ***Samtools sort***.
- We need to obtain chromosome sizes by using the Get ***Chromosome Sizes tool*** (also using hg38).
- Using the sorted BAM file, we will create a ribosome profile (BED file) file by using the ***Create Ribosome Profiles*** tool in **GWIPS-Viz (genomic alignment) branch**. The only difference here from regular Ribo-seq is that we can set offset to 12nt instead of 15nt to capture P-sites instead of A-sites.
- Next we will convert the BED file to BigWig so that it can be uploaded and visualised in GWIPS-viz by using ***Convert a BED File to a BigWig*** tool***.***
- In order to upload the resulting BigWig track file to GWIPS-viz, we will need to create a link using the Generate **Custom Track** tool. You will need to copy a link to the BigWig file and use it as input. Add name and description of the sample (INIT), as well as any chromosome position of interest, e.g. chr7:5,526,409-5,530,601 (it can be changed in genome browser). Click 'Execute'. This tool will output a file containing a link. You need to download this file and then upload to the GWIPS-viz browser.

This is how the sample (track called INIT) looks in GWIPs-viz after processing:

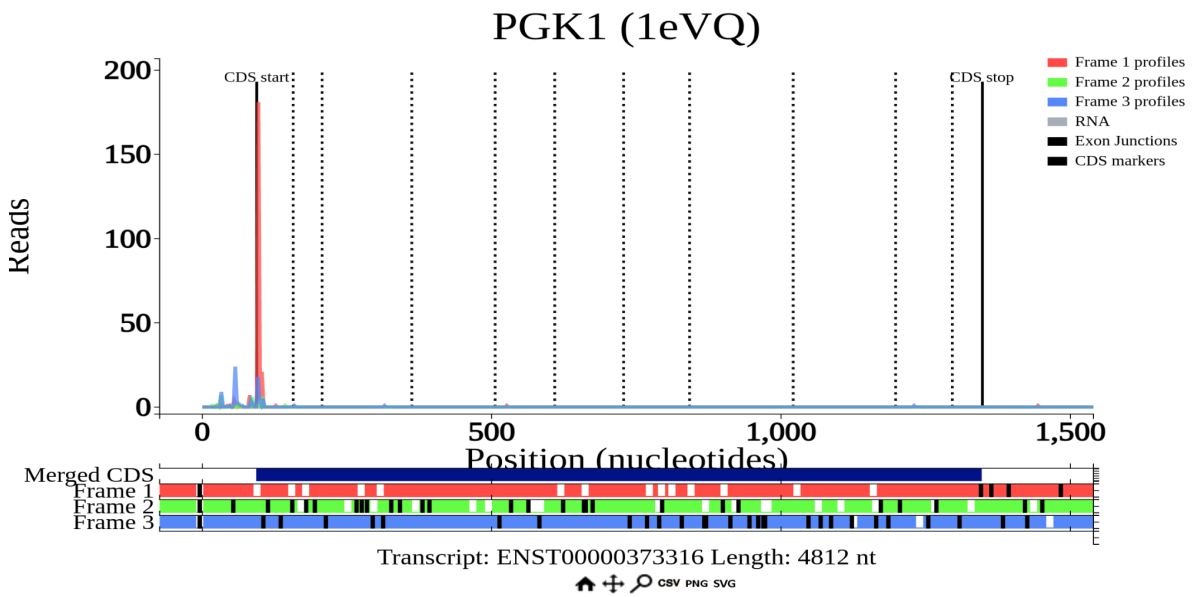


Steps for mapping to the transcriptome:

- First, we will map clean reads (footprints) to the transcriptome (GENCODE v39) using ***Bowtie Transcriptome Alignment.***
- Next, we will sort the BAM file by name using ***Samtools sort*** since this order is required for building SQLITE files. SQLITE is a format that is used to store mapped reads and associated statistics (e.g. triplet periodicity and ambiguity of mapping) for downstream analysis and visualisation in Trips-viz browser.

- Finally, we will create a SQLITE file using the ***BAM to Sqlite*** tool (GENCODE v39). We will then upload it to Trips-viz and explore TCP-seq profiles. When uploading the file to the Trips-viz, you need to select ‘Ribo-seq’ as File Type.

Here you can see subcodon Ribo-seq profiles. For *PGK1* we can see nice and clean peak at the CDS start in the correct reading frame (red):



However, for *B2M*, we see a range of peaks in the beginning of CDS, though in correct reading frame as well (green):