5′ & 3′ End-Seq Data Review Guide

Introduction

Once 5′ & 3′ End-Seq samples have been prepped and sequenced, the resulting data is processed by our expert bioinformatics scientists to generate a dataset rich in information for transcription start sites (5′ End-Seq) and/or poly-adenylation sites (3′ End-Seq). The 5′ and 3′ End-Seq data processing pipeline begins with UMI (unique molecular identifiers) trimming and adapter trimming of raw sequencing reads, then reads are filtered of repetitive genome elements such as rRNA and aligned to the reference genome (ie. human, mouse, etc.). Once aligned, PCR duplicates are removed and ends of reads are cleaned of mismatches due to the RT. Finally, single nucleotide ends are called with the PureCLIP peak calling algorithm. For 3′ End-Seq, end calls undergo a final filtering step to remove false ends derived from genomic poly-A sequence (**Figure 1**). Following data analysis, a user will receive a login and key to download several data files from our secure SFTP server, including intermediate data and detailed reports summarizing the results for each sample in the 5′ & 3′ End-Seq experiment. The 5′ & 3′ End-Seq data deliverables can be complex; to assist in understanding the rich dataset delivered, this data review guide provides a step-by-step explanation of the results, describing the different components of the figure and table in the final HTML report and each data file type delivered. For additional guidance in understanding Eclipse Bio's data deliverables, please contact techsupport@eclipsebio.com.

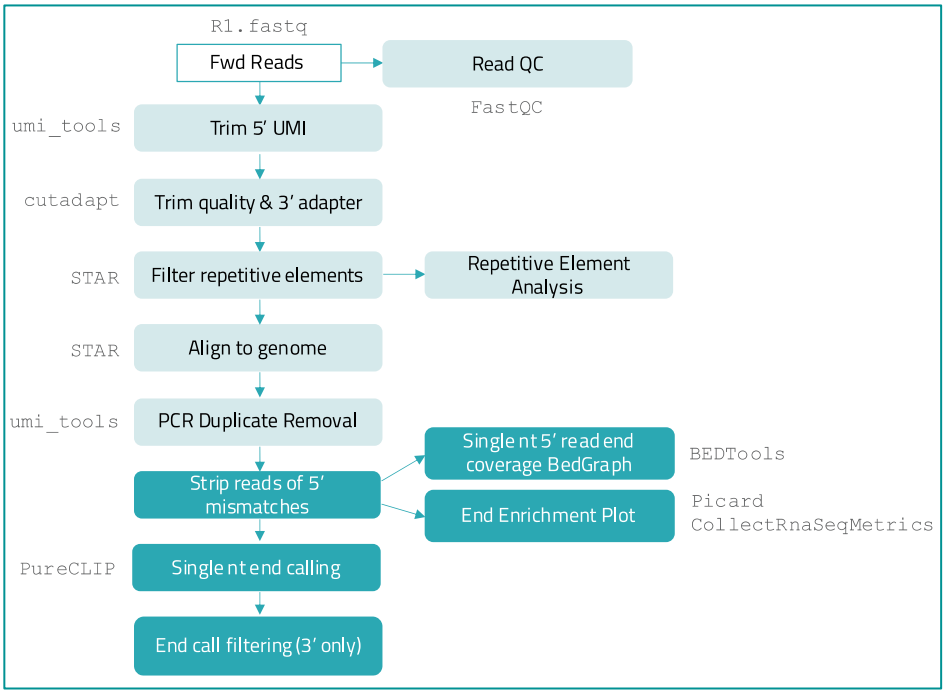


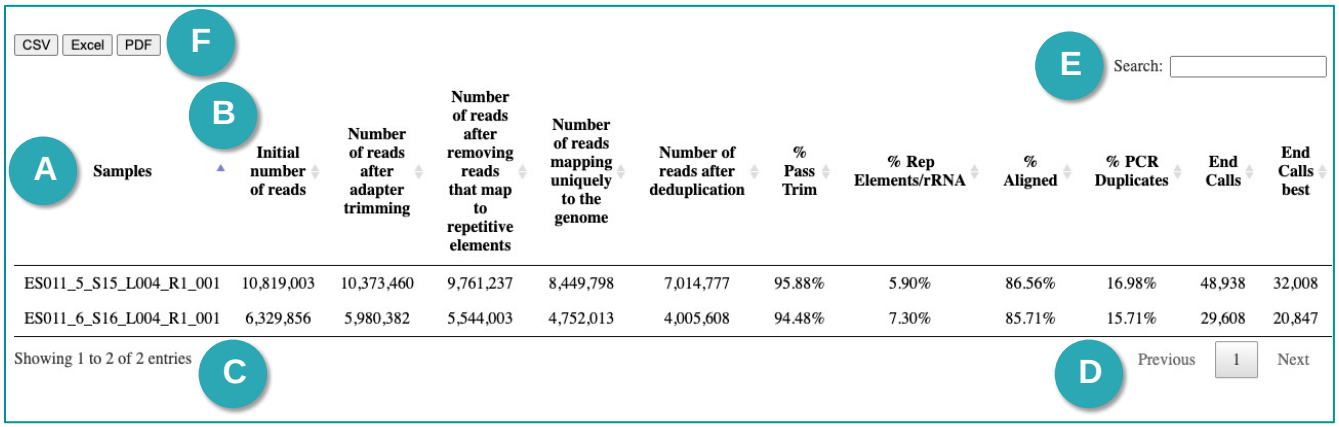
Figure 1. 5′ & 3′ End-Seq data analysis pipeline. Each analysis step is ordered from top to bottom with the publicly available tool used listed to the side of the step where available.

5′ & 3′ End-Seq HTML Report

The HTML reports summarize all the samples in the 5′ End-Seq or 3′ End-Seq experiment, providing an overview of the data alignment metrics, end calls, and end enrichment.

1. Quality Statistics table

The quality statistics table contains the read and end call metrics for each sample.



A. Column Names:

a. ***Samples*:** Sample names that are also used as part of the data file names delivered with this report. The sample name may contain "R1" to specify that the forward read (Read 1) was processed, however this is always the case whether R1 is there or not.

b. ***Initial number of reads*:** The total number of raw reads sequenced directly off the sequencer. Corresponds to the number of entries in the associated **fastq.gz** file.

c. ***Number of reads after adapter trimming*:** The result of the first filtering step, or the total number of reads remaining after removing 3′ adapter sequences. Reads that are less than 18bp after adapter trimming are too short for alignment and are removed from downstream processing.

d. ***Number of reads after removing reads that map to repetitive elements*:** The result of the second filtering step, or the total number of reads remaining after removing reads that map to our repetitive elements database (including ribosomal RNA).

e. ***Number of reads mapping uniquely to the genome*:** The total number of reads after filtering that align uniquely to the reference genome (usually human reference, hg38 unless otherwise noted).

f. ***Number of reads after deduplication*:** The total number of reads after PCR duplicate removal. Reads aligning to the same genomic location with the same 10bp UMI sequence are marked as PCR duplicates and collapsed to a single read.

g. ***% Pass Trim*:** Number of reads after adapter trimming (c) divided by Initial number of reads (b) as a percent. Reads removed at this step were mostly adapter containing.

h. ***% Rep Elements/rRNA*:** Number of reads after removing reads that map to repetitive elements (d) divided by Number of reads after adapter trimming (c) as a percent.

i. ***% Aligned*:** Number of reads mapping uniquely to the genome (e) divided by Number of reads after removing reads that map to repetitive elements (d) as a percent.

j. ***% PCR Duplicates*:** Number of reads removed during the deduplication step (f) divided by Number of reads mapping uniquely to the genome (e) as a percent.

k. ***End Calls*:** Number of single nucleotide end calls determined by the PureCLIP algorithm.

l. ***End Calls best:*** For 3′ End-Seq only, the best end calls after filtering to remove End Calls (k) that were likely derived from a nearby genomic poly-A sequence. The samples displayed above are 3′ End-Seq;

however 5′ End-Seq does not undergo this filtering and so in 5′ End-Seq reports this column is marked "NA".

B. Table Sorting: Samples are sorted alphabetically, but you may choose to sort by any column of the table. To do so, simply click the arrows in the header of the table next to this column and it will resort by this value. This example shows the default sorting.

C. Number of Entries: The table contains as many samples listed as entries at the bottom left corner, displaying up to 10 entries at a time. This example has 2 total entries.

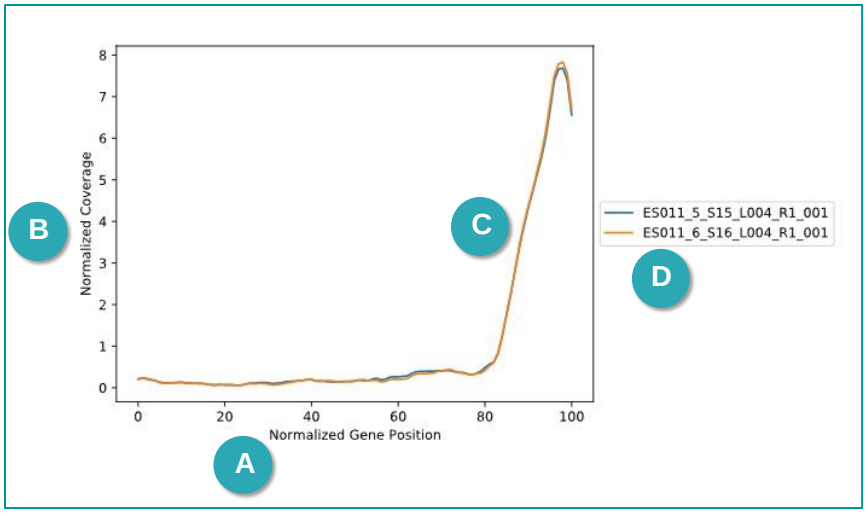
D. Page Navigator: The entries in the table can be viewed using the *Previous* and *Next* buttons on the bottom right. The next or previous 10 samples will load.

E. Search: The table is also searchable with keywords in order to list only a specific set of samples of interest. Multiple keywords can be searched with a space to separate each. This is helpful when there are many samples.

F. File Export: Selecting the buttons on the top left will export the table to a file in the selected format (CSV, Excel, or PDF). Exporting the table will preserve any sorting or filtering that has been performed.

2. Gene Coverage Plot

The gene coverage plot depicts the 5′ & 3′ End-Seq normalized read distribution across all transcripts. This plot is generated with outputs from ***picard collectRnaSeqMetrics*** using the ***.bam*** file of aligned, deduplicated End-Seq reads for each sample, and a transcript reference file (usually Gencode v35 unless otherwise noted).



A. X-axis: The normalized gene position or relative position 0%-100% across a transcript 5′ to 3′.

B. Y-axis: The normalized coverage or coverage normalized to the mean within each transcript at the relative transcript position.

C. Colored lines: The lines are each colored per sample based on the color legend on the right. As shown here, for these 3′ End-Seq samples, the read coverage peaks at the 3′ end of transcripts.

D. Legend: The list of all samples displayed in the plot and their corresponding line color.

Data File Descriptions

There are several types of data files available for download for each sample. Files will be labeled with the sample name, followed by the suffix detailed below.

***.fastq.gz**

Raw sequencing reads in FASTQ format and gzipped. This file is meant for users that want to run a data analysis pipeline beginning with raw reads. The general format of a FASTQ file contains four lines per read with the following information:

Line 1: unique sequence identifier, and may be followed by an optional description
Line 2: raw sequence
Line 3: "+", and may be followed by the unique sequence identifier and description
Line 4: sequence quality values

***.adapterTrim.round2.fastq.gz**

Reads with UMI relocated to the read name and trimmed of adapter at the 3′ end, also in FASTQ format and gzipped. This file is meant for those who want to visualize cleaned reads and run a data analysis pipeline beginning with trimmed reads. The general format of a FASTQ file is detailed above (for **.fastq.gz**).

***.adapterTrim.round2.rmRep.sorted.rmDup.sorted.bam**

Reads that have been filtered of repetitive elements, aligned to the reference genome, and removed of PCR duplicates. BAM files are compressed binary versions of Sequence Alignment/Map (SAM) files. BAM files are used as input to downstream data analysis and for visualizing read alignments in a genome browser (Note: the corresponding ***bam.bai** index file is also required for loading BAM files into a genome browser). To explore the contents of a BAM file, it will need to be uncompressed back to SAM format using the command line tool `samtools` (available through http://samtools.sourceforge.net/). A SAM file begins with a header detailing the reference genome and is followed by one read alignment per line. The general format for each alignment in a SAM file is as follows:

Column 1: query template name (ie: read name)
Column 2: bitwise flag (see https://broadinstitute.github.io/picard/explain-flags.html for a guide to SAM flags)
Column 3: reference sequence name (ie: chromosome for alignment)
Column 4: 1-based leftmost mapping position
Column 5: mapping quality
Column 6: Cigar string
Column 7: reference name of the mate/ next read
Column 8: position of the mate/ next read
Column 9: observed template length
Column 10: segment sequence
Column 11: ASCII of Phred-scaled base quality+33

***adapterTrim.round2.rmRep.sorted.rmDup.sorted.bam.bai**

The index file associated for the above BAM file. BAM index files are typically required when interacting with BAM files either via command line or in a genome browser. BAM index files should be stored in the same folder as BAM files so that tools can locate them.

***snt_[neg/pos].bedgraph**

The single nucleotide 5′ end read coverage bedgraph files. The 5′ end of the read will start at the transcription start site (TSS) of a transcript in the case of 5′ End-Seq or the poly-adenylation site (PAS) in the case of 3′ End-Seq. Coverage is only counted at the 5′ most single nucleotide position of the read to show where saturations TSS or PAS sites occur. This bedgraph file can be used to display continuous data in a genome browser. Files ending in ***snt_pos.bedgraph** are only the counts on the "+" strand, while files ending in *** snt_neg.bedgraph** are only the counts on the "–" strand.

***ends.bed**

Locations of single nucleotide end calls equivalent to PAS (3′ End-Seq) or TSS (5′ End-Seq) identified with PureCLIP (https://github.com/skrakau/PureCLIP). The BED file contains the following tab-separated columns:

Column 1: Chromosome
Column 2: Start position
Column 3: Stop position
Column 4: PureCLIP state score
Column 5: PureCLIP probability score
Column 6: Strand (+ or -)
Column 7: PureCLIP data string

This file can be loaded into a genome browser for visualization and manipulated with command line tools such as BEDTools (https://bedtools.readthedocs.io/). PureCLIP aims at detecting single nucleotide positions where a significant fraction of read starts accumulate at the same position forming a 'read cliff', originating from a gene end.

***ends.best.bed** - *This file is only included for 3′ End-Seq samples.*

Locations of the best end calls or PAS from a 3′ End-Seq experiment. These ends have been filtered for false positives derived from genomic poly-A sequence. The BED file contains the following tab-separated columns:

Column 1: Chromosome
Column 2: Start position
Column 3: Stop position
Column 4: Name
Column 5: PureCLIP probability score
Column 6: Strand (+ or -)

This BED file can be loaded into a genome browser for visualization and manipulated with command line tools such as BEDTools (https://bedtools.readthedocs.io/).

***summary_report.html**

Report with tables and enrichment. See the above section for a detailed review of each aspect of the HTML report.