At Arbor, we are using various screening methods to interrogate the functionality of large numbers of gene systems. As part of this process, our recent characterization of Cas13d involved characterizing the ability of this enzyme to truncate CRISPR pre-crRNAs to form crRNAs, which guide the enzyme to bind and cleave specific RNA targets. In the below problem, you will write a basic pipeline to process RNAseq data to evaluate the Cas13d crRNA processing.

Download data

1. Download the following files from NCBI SRA:
   - download NCBI SRA Toolkit from: https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software
   - download paired end fastq RNAseq data for SRR6800317 (2 Es and 2 Rsp CRISPR pre-crRNAs treated inVitro with Es or RspCas13d, respectively)
     - `fastq-dump -I --split-files SRR6800317`
     - Read 1: SRR6800317_1.fastq; Read 2: SRR6800317_2.fastq
   - download paired end fastq RNAseq data for SRR6800318 (Es or Rsp CRISPR pre-crRNAs – no Cas13d treatment)

Format and align fastq data (python preferred)

2. Use Biopython SeqIO.parse (format = 'fastq') to read the records from each pair of fastq files, and check that read1 (R1) and read2 (R2) files are sorted correctly (sequence names should match for the read pair at each index in R1 and R2 files.

Note: Normally R1 and R2 reads have the same name in Illumina fastq files, allowing direct matching. However NCBI SRA renames reads in submitted fastq files, so matching must be done using only a substring within the R1 and R2 read names (highlighted in blue below).

Example R1 fastq read

```
@SRR6800317.1.1 1 length=75
CAACTNACTGAAATGCGAACTACACCCGTGCAAAATTGCAGGGGTCTAAATCAGATCGGAAGAGCACACGTCTGA
+SRR6800317.1.1 1 length=75
AAAAA#EEEE/<EEEEAEEE<AEEAEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEAAEAEEAAEEAEEEE/EEEE
```

3. Trim R1 and R2 reads to the first 30bp (the start/end of the original RNA molecule).

Note: 30bp is long enough providing high mapping fidelity while minimizing failed mappings due to accumulated sequence errors introduced by library prep or sequencing in long reads.

4. Discard all 30-mer reads with average Illumina Phred quality scores less than 30. Write truncated reads to new R1 and R2 fastq files.

5. Build a bowtie2 index for the file ‘pre-crRNA_reference.fa’ using default parameters.
   - for documentation on bowtie2, visit: http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml

6. Align each pair of truncated fastq files to ‘pre-crRNA_reference.fa’ index using bowtie2.
   - Contains 2 Es and 2 Rsp CRISPR pre-crRNA reference sequences

Analyze RNAseq alignments (python preferred)

7. Install samtools python package

- pip install pysam
- Documentation: https://media.readthedocs.org/pdf/pysam/latest/pysam.pdf

8. Read alignment sam file using ‘pysam.AlignmentFile’, and for each read pair in the sam file, determine the following the R1 and R2 reads both align to a single reference.

9. For instances where both R1 and R2 are aligned, determine:
   - Determine the start, end, and orientation of the R1 and R2 reads on the reference.
   - Reconstruct the sequence of the original RNA molecule given the reference.

10. Calculate a frequency distribution of different truncated Es and Rsp pre-crRNAs reconstructed in step 7, and determine truncated sequences for each pre-crRNA that are highly enriched in samples treated with Cas13d vs. those with no Cas13d treatment.

Does the data suggest that Cas13d is manipulating the pre-crRNA reference sequences, and if so, what are some of the most prevalent truncated forms (possible crRNAs)? While the scope of this problem is small (4 arrays treated with/without 2 genes), in practice, what are some ways of scaling this analysis for 100s or 1000s of gene-crRNA combinations?