MA-GenTA Analysis - Data Processing

Jacquelynn Benjamino

August 3, 2020

This code entails the pipeline for analysis of raw reads from the MA-GenTA assay. Sample names and file names are designated by SAMPLE_NAME and FILENAME.

Open-source software used in this analysis

```
Cutadapt v.1.14
Kraken2 v.2.0.8-beta
BWA v.0.7.12
SAMtools v.0.1.19
USEARCH
BEDtools v.2.27
```

Build an index from the reference genomes for mapping

```
bwa index -a bwtsw REFERENCE_GENOMES.fa
```

Run BWA mapping

```
bwa mem -t 8 -M REFERENCE.fa RAW_FASTQ_FILENAME.fastq.gz | samtools view -Shb - -o FILENAME_bwa.bam
```

Sort the bam file

```
samtools sort FILENAME_bwa.bam -o FILENAME_bwa_sort.bam
```

Remove alignments with a bitscore of 256 (secondary alignment)

```
samtools view -F 256 -bh FILENAME_bwa_sort.bam > FILENAME_bwa_primary.bam
```

Remove the reads that mapped to multiple sites and keep only unique mapped reads

```
samtools view -F 4 FILENAME_bwa_primary.bam | grep -v XA: | grep -v SA: > FILENAME_unique.txt
```

Create a file of unique hits to use for BEDtools

Grep the header from the sorted bam file

```
samtools view -H FILENAME_bwa_primary.bam > FILENAME_header.txt
```

Combine the header with the unique.txt file to create a bam file of unique hits

```
cat FILENAME_header.txt FILENAME_unique.txt | samtools view -b FILENAME_unique.bam
```

Run bedtools intersect to match the probes to the bins for each read. The BED file (PROBE_LOCATIONS_BEDFILE.bed) used depends on V2 (Allegro) or V4 (JAX) probe design

```
bedtools intersect -c -a PROBE_LOCATIONS_BEDFILE.bed -b FILENAME_unique.bam > FILENAME_probe_counts.txt
```

Rename probe count columns

```
echo $'Bin\tstart\tend\tProbe\tlength\tstrand\tSAMPLE_NAME' | cat - FILENAME_probe_counts.txt > FILENAME_pro
be_counts_named.txt
```

Keep only the Bin, Probe, and count columns

```
cat FILENAME_probe_counts_named.txt | cut -f1,4,7 > FILENAME_probe_counts_named_cut.txt
```

Python script to combine the files (This is in a python script that gets run through a shell script)

```
# Make a dataframe for each file (1-n)
df1-n = pd.read_table('FILENAME_probe_counts_named_cut.txt', sep='\t')

# Combine data frames for V2 (Allegro) set and separate for V4 (JAX) set
data_framesV2 = [df4, df99, df100, df101, df102, df103, df104, df105, df106, df107, df108, df109, df110, df
111, df112, df113, df114, df115, df116, df117, df118, df119, df120, df121, df122, df123, df124, df125, df12
6, df127, df128, df129, df130, df131, df132, df133, df134, df135, df136, df137, df138, df139, df140, df141,
df142, df143, df144, df145, df2, df5, df6, df7, df8, df9, df10, df11, df12, df13, df14, df15, df16, df17, df
18, df19, df20, df21, df22, df23, df24, df25, df26, df27, df28, df29, df30, df31, df32, df33, df34, df35, df
36, df37, df38, df39, df40, df41, df42, df43, df44, df45, df46, df47, df48, df49, df50, df51]

# Concatenate the dataframe
df_merged_V2= pd.concat(data_framesV2, axis=1)

# Print the dataframe to a csv file
pd.DataFrame.to_csv(df_merged_V2, 'mergedV2.txt', sep='\t', na_rep='.', index=False)

# In excel, make sure rows are all in the same order and then remove the bin, probe columns for all samples
to make a counts table
```

Get mapping stats

Get total reads after 97.5/50, rename, and concatenate them

```
samtools view -c FILENAME_bwa_sort.bam -o FILENAME_total_reads.txt
    f=$(ls -1tr FILENAME_total_reads.txt | head -1); echo $f | cat - $f > FILENAME_total_reads_named.txt

cat * total_reads_named.txt > total_reads_combined.txt
```

Get mapped reads, rename, and concatenate them

```
samtools view -c -F 4 FILENAME_bwa_primary.bam -o FILENAME_mapped_reads.txt
    f=$(ls -1tr FILENAME_mapped_reads.txt | head -1); echo $f | cat - $f > FILENAME_mapped_reads_named.txt

cat *mapped_reads_named.txt > combined_mapped_reads.txt
```

Get unique reads, rename, and concatenate them

```
samtools view -c FILENAME_unique.bam -o FILENAME_unique_reads.txt
    f=$(ls -1tr FILENAME_unique_reads.txt | head -1); echo $f | cat - $f > FILENAME_unique_reads_named.txt

cat *unique_reads_named.txt
```

Get on target mapping value, rename, and concatenate them

```
bedtools intersect -abam FILENAME_unique.bam -b target_snps.bed -u -bed | wc -l > FILENAME_on_target.txt
    f=$(ls -1tr FILENAME_on_target.txt | head -1); echo $f | cat - $f > FILENAME_on_target_named.txt

cat *on_target_named.txt > combined_on_target.txt
```