End Motif Tutorial

Jeffrey Szymanski

August 23, 2024

Notes and Dependencies

- This is a minimal working example intended to be run from the main repo directory
- sample_motifs.py is currently designed ONLY for use with BWA- other aligners will have different samtools flags called.
- R and python packages as specified per script.
- Small bam files human1-5 in data/bams
- Appropriate reference fasta, `GCA_000001405.15_GRCh38_no_alt_analysis_set.fna` (available from ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.15_GRCh38/seqs_for_alignment_pipelines.ucsc_ids/GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.gz)

1. Clear motifs data directory if present:

```
if [ -d data/motifs ]; then rm -rf data/motifs; fi
mkdir data/motifs
```

2. Generate motifs from 5' ends of bams:

```
scripts/sample_motifs.py -h

REFERENCE_GENOME="/home/jeszyman/pnst/inputs/GCA_000001405.15_GRCh38_no_alt_analysis_se
↪  t.fna"

scripts/sample_motifs.py --bam_file data/bams/human1.bam \
                         --output_file data/motifs/human1_motifs.tsv \
                         --motif_length 4 \
                         --num_reads 10000 \
                         --threads 4 \
                         --reference_genome $REFERENCE_GENOME

for bam in data/bams/*.bam; do
    base_name=$(basename "$bam" .bam)

    output_file="data/motifs/${base_name}_motifs.tsv"

    scripts/sample_motifs.py --bam_file "$bam" --output_file "$output_file" \
```

```
                                --motif_length 4 --num_reads 10000 \
                                --threads 4 --reference_genome "$REFERENCE_GENOME"
done
```

3. Consolidate to a single matrix:

(In R)

```r
library(tidyverse)

# Function to read a file and format it for merging with counts
read_motif_file_counts <- function(file) {
  df <- read_tsv(file, col_names = c("motif", "count"))
  file_name <- tools::file_path_sans_ext(basename(file))
  df <- df %>% rename(!!file_name := count)
  return(df)
}

# Function to read a file and format it for merging with fractions
read_motif_file_fractions <- function(file) {
  df <- read_tsv(file, col_names = c("motif", "count"))
  total_count <- sum(df$count)
  df <- df %>% mutate(fraction = count / total_count)
  file_name <- tools::file_path_sans_ext(basename(file))
  df <- df %>% select(motif, fraction) %>% rename(!!file_name := fraction)
  return(df)
}

# List of files
files <- list.files(path = "data/motifs", pattern = "*.tsv", full.names = TRUE)

# Read and merge all files for counts
motif_data_counts <- files %>%
  map(read_motif_file_counts) %>%
  reduce(full_join, by = "motif")

# Read and merge all files for fractions
motif_data_fractions <- files %>%
  map(read_motif_file_fractions) %>%
  reduce(full_join, by = "motif")

# Save the resulting matrices
write_tsv(motif_data_counts, "data/combined_motif_counts_matrix.tsv")
```

```
write_tsv(motif_data_fractions, "data/combined_motif_fractions_matrix.tsv")
```

4. Motif diversity score

(In R)

```
annotation = data.frame(library =
↪  c("human1_motifs","human2_motifs","human3_motifs","human4_motifs","human5_motifs"),
                        cohort = c("healthy","cancer","healthy","cancer","healthy"))
annotation

motifs = read_tsv("data/combined_motif_fractions_matrix.tsv")

motifs

motifs_long <-
  pivot_longer(motifs, cols = !motif, names_to = "library", values_to = "fraction") %>%
  left_join(annotation, by = "library") %>%
  select(motif, library, fraction, cohort)

motifs_long

mds = motifs_long %>%
  mutate(fraction = ifelse(fraction == 0, 1e-10, fraction)) %>%  # Add a small constant
↪  to avoid log(0)
  mutate(mds = -fraction * log(fraction) / log(256)) %>%
  group_by(library) %>%
  summarize(mds = sum(mds))

mds
```

5. F-profiles

```
scripts/fprofiles.py -h

scripts/fprofiles.py --data_file ./data/combined_motif_fractions_matrix.tsv
↪  --output_fprof_per_lib ./data/fprofiles.tsv --output_motif_per_fprof
↪  ./data/motif_per_fprofile.tsv
```