Lab Assignment 3: Sequence Comparison

2022-02-04
Part 1: Finding the difference between hg38 and CHM13
Part 1: Compare hg38 to CHM13

- hg38 is a chimeric reference from a few individuals
  - Maintained and developed by the Genome Research Consortium
  - Many gaps and satellite contigs
- CHM13 is from the CHM13 cell line that expresses human telomerase reverse transcriptase (hTERT) (retains 46,XX karyotype)
  - Developed by the T2T consortium
Part 1: Compare hg38 to CHM13

- Copy the finding_gaps_template.py script to your directory
  - /home/assignments/assignment3/finding_gaps_template.py

- Complete the script to do the following:
  - Count nucleotide frequencies
  - Count dinucleotide frequencies
  - Count and gaps and gap sizes
  - Plot gap size distribution
Part 2: Align Reads to Chr22 and mark Duplicates
Step 1: Create Indices for hg38 and CHM13

- Create an index for both assemblies of chr22
- Example command in assignment

```bash
asouthard-smith@genomics:$ bowtie2-build
No input sequence or sequence file specified!
Bowtie 2 version 2.3.4.1 by Ben Langmead (langmea@cs.jhu.edu, www.cs.jhu.edu/~langmea)
Usage: bowtie2-build [options]* <reference_in> <bt2_index_base>
    reference_in        comma-separated list of files with ref sequences
    bt2_index_base      write bt2 data to files with this dir/basename
*** Bowtie 2 indexes work only with v2 (not v1). Likewise for v1 indexes. ***
Options:
  -f               reference files are Fasta (default)
  -c               reference sequences given on cmd line (as
                   <reference_in>)
  --large-index    force generated index to be 'large', even if ref
                   has fewer than 4 billion nucleotides
  -a/--noauto      disable automatic -p/--bmax/--dcv memory-fitting
  -p/--packed       use packed strings internally; slower, less memory
  --bmax <int>      max bucket sz for blockwise suffix-array builder
  --bmaxdivn <int>  max bucket sz as divisor of ref len (default: 4)
  --dcv <int>       diff-cover period for blockwise (default: 1024)
  --noded           disable diff-cover (algorithm becomes quadratic)
  -r/--noref       don't build .3/.4 index files
  -3/--justref     just build .3/.4 index files
  -o/--offrate <int> SA is sampled every 2^<int> BWT chars (default: 5)
  -t/--ftabchars <int> # of chars consumed in initial lookup (default: 10)
  --threads <int>   # of threads
  --seed <int>      seed for random number generator
  -q/--quiet        verbose output (for debugging)
  -h/--help         print detailed description of tool and its options
  --usage           print this usage message
  --version         print version information and quit
```
Step 2: Align with bowtie2

- Reads are in:
  
  `/home/assignments/assignment3/test-500k.fq`

- The fastq file is unpaired

- Save both the alignment and report output files

Step 3: Remove Duplicates with Samtools

- Duplicate reads are measured from the position of the 5’ end of the read.
- Duplicates are the result of non-uniform amplification and hybridization rates during library construction & sequencing.


http://www.htslib.org/algorithms/duplicate.html
Step 3: Remove Duplicates with Samtools

- 4 stages:
  - Collate
  - Fixmate
  - Sort
  - MarkDuplicates

- markdup output should be saved to a file

- markdup report will be output to terminal
Part 3: Running BLAST
Step 1: Obtain Gene Sequence

- Obtain **coding** sequence for **RAP1** from [yeastgenome.org](http://yeastgenome.org)
Step 1: Obtain Gene Sequence
Step 2: Run BLASTx Web Tool

- Compare BLOSUM62 vs. BLOSUM80
- Compare default existence (11), extension (1) penalties vs existence of 7, extension of 2
Step 2: Run BLASTx Web Tool
Step 2: Run BLASTx Web Tool
Things to turn in:

• README.txt
• finding_gaps.py script
• hg38-chr22_gap_distribution.png
• Bowtie2 Alignment files
• Bowtie2 Report files
• Samtools duplicates removed sam file
• Extra credit items:
  • Additional gap size distribution plot
  • Python script for tallying duplicates from the bowtie2 aligned sam file.