Assignment 2: BLAST (OFF!)

Genomics

1/29/16

What to do: Assignment 2

- Provided:
  - Assignment 2
  - Reads.fq
  - Nuc_count_final.py
  - README.txt
  - Unknown.fsa (extra credit)
- To submit:
  - Completed README.txt
  - Bowtie2 output file
  - Bowtie2 report file
  - Extra credit BLAST output file

Outline of Assignment 2

- Part 1:
  - Use online BLAST to search a protein database for homology to given gene. Adjust parameters (BLOSUM) and observe difference in results.
- Part 2:
  - Use bowtie2 to index human chromosome 22
  - Use bowtie2 to align reads to chr22.
    - Specify your alignment parameters to give you desired output.
  - Count the frequency of single and di-nucleotides in reads.fq and identify if enrichment of certain elements is present.
    - Determine what experimental assay that the reads.fq could have originated from.

Basic Local Alignment Search Tool

- Algorithm using hash tables (next lecture!) to align sequences of proteins or nucleotides.
- Multiple databases available depending on query type:

	Database	Query
blastn	Nucleotide	Nucleotide
blastp	Protein	Protein
blastx	Protein	Translated nucleotide
tblastn	Translated nucleotide	Protein
tblastx	Translated nucleotide	Translated nucleotide

- GenBank – primarily uncurated sequence from ~250,000 named organisms
- PBD (Protein Database Bank) – 3D protein structures
- RefSeq- Curated single examples of model organism DNA, RNA, and protein
- nr (non-redundant) – filter out duplicate entries from databases mentioned above

Fastq format

FORMAT:

@SEQ_ID (Read name/identifier)

GATCATGCATGCATGCTAGCTGATCTAGCTATGCTAG (Sequence)

+ (Strand info)

!’’’*(())(*&&%%%()%*%*%.13.,.1,3.,5CCF>>CADE (ASCII scores)

ASCII (American Standard Code for Information Interchange) quality scores:

Increasing quality

!"#$%&'()*+,-./0123456789:;<=>@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~

Quality Score: The probability that the corresponding base call is incorrect.

Command-line BLAST

- BLAST module has been installed in our server!
- Benefits of command-line BLAST:
  - 1) Much faster than online version
  - 2) Higher throughput. List of sequences can be uploaded to be aligned
  - 3) Useful adjustable parameters can be set to filter results
    - gapopen or gapextend
    - Evalue
    - Wordsize
    - BLOSUM62 vs BLOSUM80
- Extra Credit opportunity!
  - Please use command-line BLAST to align and identify what organism the sequence (unknown.fsa) originates from.
  - Commands for BLAST can be easily found online. Please use right parameters for this exercise.

Comparing dinucleotides

- Use nuc_count_final.py to calculate frequencies of each nucleotide and dinucleotide.
- Food for thought: What is the definition of enrichment?
  - Compare chr22 nucleotide and dinucleotide frequencies to those in reads.fq dataset.
- In the README.txt:
  - Justify which nucleotide/dinucleotide frequencies have relevant differences
  - What type of experiments can generate this data?
    - Explain why, biologically, that experiment would produce this pattern.

Questions from Assignment 1?

- Let us know!