Short Read Alignment

The read mapping problem

- Next generation sequencing (NGS) creates short reads
- These must be mapped to one or more reference genomes (billions of bp)
- These reads may have errors in them
- The orientation of the read relative to the genome is not known
- We may not have the exact reference genome/ genomes for these reads
- How can we align the reads to a reference accounting for inexact matching and also within a reasonable amount of time and memory space?

Short Read Alignment Uses

- May wish to assemble reads into a genome (not covered here)
- May wish to align reads to one reference genome for genetic variation analysis
- May wish to align a microbiome to a set of reference genomes for species or functional analysis

Strategies

- Reference indexing:
  - Reference likely to have repetitive sections
  - Split the reference into given words
  - Often use hash tables to allow for rapid searching
  - Indexing can be done on query sets too in a similar manner
- Spaced seed indexing
  - Split reads into segments
  - Align segments in pairs to the references (also split into seeds)
    - All segments align, perfect alignment
    - 1 segment doesn't align, 1 mismatch
- Burrows-Wheeler
  - Indexes reference using the B-W transformation
  - Reads split into suffixes, each 1 character shorter than the previous
  - Match suffixes to index, increasing one nucleotide at a time
  - If perfect alignment is found, read is aligned
  - If perfect alignment not found, change 1 character and realign

Single Reference Genome

- Can use SRA for SNP analysis, splice variants
- Often use a local alignment strategy for placing the reads onto a reference genome
- Some software listed here: http://www.cbcb.umd.edu/research/SR-assembly.shtml
- BWA: Burrows-Wheeler Aligner:
  - http://bio-bwa.sourceforge.net/
  - Allows for gapped alignment and mismatches
  - Outputs files useable by SNP callers (SAM files)
- BowTie
  - http://bowtie-bio.sourceforge.net/
  - Also uses the BW algorithm
  - Often used for human genome alignment
  - Used by many other programs such as TopHat which aligns RNA-Seq data
- SOAP: Short Oligonucleotide Analysis Package
  - http://soap.genomics.org.cn/
  - Uses a Burrows-Wheeler transformation with a seed strategy
  - Has options for SRA, SNP calling, RNA-Seq analysis
- Segmenator
  - http://bioinf.man.ac.uk/robertson/segminator/
  - For aligning reads to viral genomes
  - Can perform platform error correction and phylogenetic inference

Multiple Genomes

- Often part of microbiome analysis
- First step in assessing species abundance or functional assignments
- Requires a template alignment
- PyNAST: Python Nearest Alignment Space Termination
  - http://qiime.org/pynast/
  - Uses a reference alignment, BLAST and gap realignment
  - Most often used with a 16S dataset
- Mothur:
  - http://www.mothur.org/wiki/Main_Page
  - Uses N-W, Gotoh or blastn algorithm with gap refinement using NAST