BST281: Genomic Data Manipulation, Spring 2019

Wednesday 02: Sequence assembly, annotation, and analysis

Quality control.

Quality trimming, length filtering, deduplication.

Assembly.

de Bruijn graphs: counts of all overlapping k-mers for some word length k across all reads.

Remove tips and small bubbles (errors), retain large cycles (ambiguous repeats).

Evaluation by N50/NG50: contig length *L* for which 50% of all bases are in contigs of length < *L*.

Methods continue to improve, but remain highly memory-intensive.

Annotation.

Open reading frame (ORF) calling, often by Hidden Markov Models (HMMs).

Find start-like regions, followed by codon-like regions, followed by stop-like region.

QC for length, strand, up/down-stream organization, secondary structure.

Functional annotation: also by HMMs and/or annotation transfer by homology.

Easy to approximate, very difficult to finish.

Mapping / accelerated search.

Continuum of options available from exact-slow to approximate-fast alignment.

Sequencing applications.

Variant and peak calling, contact mapping, RNA-seq, competitive growth, Tn-seq, etc.

Lots of file formats and tools.

FASTQ, SAM/BAM, BAI, pileup, GFF, BED, VCF/BCF, WIG...

SAMtools, FastX/QC, Trimmomatic, AMOS, HMMer, USEARCH, DIAMOND, bowtie2, BWA, GATK...

# Textbooks

Assembly: Pevsner, Chapter 4 p121-160

Annotation: Pevsner, Chapter 15 p700-737

Applications: Pevsner, Chapter 17 p797-830

Genomes: Lesk, Chapter 7 p215-233

# Literature

[Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. Bradnam et al, Gigascience 2013](https://www.ncbi.nlm.nih.gov/pubmed/23870653)

[Evaluation of next-generation sequencing software in mapping and assembly. Bao et al, J. Hum. Gen. 2011](https://www.ncbi.nlm.nih.gov/pubmed/21525877)

[Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Dekker et al, NRG 2013](https://www.ncbi.nlm.nih.gov/pubmed/23657480)