BST281: Genomic Data Manipulation, Spring 2017

Wednesday 04: Sequence assembly, annotation, and analysis

Sign up for journal club!

Email instructional team with group, paper, date, and justification by end of today.

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*.

Annotation.

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

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

Mapping / accelerated search.

Accelerated homology search: heuristics that limit the number or minimum homology of aligned sequences.

Mapping: very rapid high-identity matching of short reads to best hit(s) in a target sequence.

Often using the Burrows-Wheeler Transform for efficient indexing.

Sequencing applications.

Variant and peak calling, contact mapping, RNA-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

Mapping: Pevsner, Chapter 5 p167-198

Annotation: Pevsner, Chapter 15 p700-737

Applications: Pevsner, Chapter 17 p797-830

# 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)