The first midterm (Thursday Oct 10 in-class) covers in-class material days 1-8, labs 1-4, reading weeks 1-4. You may bring a 1 page (front and back), hand-written study guide, but no other notes or resources. You will not need a calculator. I have put vocab in blue.

1. Pattern Matching and Boyer-Moore
   - String terminology and types of string search problems
   - Notation (including m length of pattern P and n length of search string S)
   - Naive string search algorithm and runtime
   - Boyer-Moore string search algorithm and runtime
   - Details of Boyer-Moore: bad character table, good suffix table etc
   - k-mer hashing: general idea, not details
2. BWT and Read Mapping
   - What is read mapping? What is the input; what is the output?
   - What is the Burrows-Wheeler Transform (BWT) of a string S? Why was it originally used?
   - How can we recover the original string from the BWT? *Why* does this process work?
   - How much time and space does it take to construct the BWT?
   - FM-Index (BWT/L + occ + M), plus additional data structures F and A (suffix array)
   - How can we use the FM-Index for exact pattern matching (i.e. the recursive formulas for start point and end point in F)? *Why* does this work?
   - How do we use the suffix array A to find the pattern locations in the original string?
   - Common variables: n = length of genome, m = length of each read, R = number of reads
   - What are the time and space requirements of error-free read mapping? (in terms of n, m, R)
   - High-level idea of how BWA and Bowtie deal with mismatches (errors and variation)
3. Genome Assembly
   - High-level next-generation sequencing (NGS) process (obtain short reads, not entire genome)
   - What is the goal of genome assembly? What is the input; what is the output?
   - Vocab: long read, short read, base pair (bp), coverage (+ how to compute coverage)
   - Overlap graph assembly (often called Overlap Layout Consensus (OLC) assembly)
   - How do we detect overlaps between reads? How do we build the overlap graph? What would an ideal overlap graph look like? How can we simplify the overlap graph?
   - What is the runtime of building an overlap graph and why is it prohibitive?
   - What affect do sequencing errors and repeats have on graph-based genome assemblers?
   - De Bruijn Graph (DBG) assembly: how to build and traverse a DBG to create contigs
   - What is a k-mer and how should we choose it relative to m?

- Additional vocab: directed multigraph, in-degree, out-degree, balanced, semi-balanced, Eulerian path/cycle, connected component
- Traversal algorithms: Fleury's algorithm and its recursive implementation
- Time and space requirements of building and traversing a DBG
- High-level idea (not all the details) of the modifications Velvet uses to make DBGs practical
- Assembly evaluation: both by N50 and pairwise sequence alignment (if ground truth known)

4. Pairwise Sequence Alignment

- What is the goal of sequence alignment? What is the input; what is the output?
- What is the difference between local and global alignment?
- Vocab: dynamic programming (DP), homologous, substitution, gap: insertion or deletion
- Constructing and filling in a dynamic programming table, back-tracing to find the alignment
- Modifications for global (Needleman-Wunsch) vs. local (Smith-Waterman) alignment
- Three types of sequences in molecular biology: DNA (A,C,G,T), RNA (A,C,G,U), and Protein (amino acids)
- How do we weight gaps, matches, mismatches? (BLOSUM matrix for proteins)
- Do *not* need to memorize map from codons to amino acids (including start/stop)
- Multiple ways to trace back from a given cell vs. multiple cells with max score (local only)
- Modifications to the DP algorithm to produce overlap/containment alignments
- Runtime of Needleman-Wunsch and Smith-Waterman in terms of sequence lengths

5. Multiple Sequence Alignment and Phylogenetics Intro

- What is the multiple sequence alignment (MSA) problem? Sum of pairs method for scoring
- Basic idea of K-dimensional dynamic programming and its runtime
- Alternative of progressive alignment and CLUSTAL-W method
- Relationship between multiple sequence alignment and evolutionary tree
- What does genetic variation represent? Evolutionary process of mutations on tree branches
- Genetic variation vocabulary (see Handout 8)