CS 394C
Questions about Assemblers

1. What is a "read"?
2. What is meant by "3-fold coverage"?
3. What is a "DNA fragment"?
4. What is a "mate-pair"?
5. What is the meaning of "double-barreled shotgun sequencing?"
6. Describe the properties of Sanger sequencing data, and explain why it was called "Sanger sequencing".
7. What is meant by "next generation sequencing technologies" (NGS) and what is their major advantage over Sanger sequencing?
8. What is meant by a "contig"?
9. Describe basic differences in read lengths produced by Sanger sequencing, 454 sequencing, and Illumina sequencing.
10. What is the length of a typical contig in a Sanger assembly of a large genome? What about in a 454 assembly? What about in an Illumina assembly?
11. What is a typical "coverage" for a short-read assembly? What impact does this have on assembly algorithm design?
12. What is a bubble? What causes it, and how is it handled?
13. What is the major use of mate pairs in genome assembly?
14. Which of the standard sequencing technologies typically are used with mate pairs?
15. What is meant by a "homopolymer region"? Which sequencing technology has trouble with this, and what happens?
16. What is the major error issue with Helicos sequencing technologies?
17. What is a "phred quality value"?
18. What is meant by "*de novo* assembly"?
19. What is meant by "comparative assembly"?
20. Describe a basic technique for comparative assembly.
21. What complications cause comparative assembly to be challenging?

22. What is meant by "genome assembly finishing"?

23. What is meant by "re-sequencing"?

24. Compare challenges in assembly of bacterial genomes and a human genome.

25. Are assemblies always based on just one type of sequencing data, or are some assemblies based on two or more types of sequencing data?

26. Describe what an "Overlap-Layout-Consensus" (OLC) assembly technique does.

27. To avoid the computation of overlaps between all pairs of reads using an OLC assembly technique, an "indexing strategy" can be used. Give an example of one.

28. Describe a "greedy assembly" technique, and give an example where it can make a mistake.

29. What is meant by a "unitig" and a "fork" (in the context of the Celera OLC assembler)?

30. What applications are assemblies that only output unitigs sufficient for?

31. What is meant by the "k-mer spectrum"?

32. How does the Eulerian Assembly technique work?

33. What are some advantages of the de Bruijn graph (using the Eulerian assembly technique) over the OLC technique?

34. What are some challenges in using the de Bruijn graph?

35. Name two Eulerian assemblers (i.e., that use de Bruijn graphs)

36. Name some OLC assemblers.

37. Describe the technique used by the Newbler assembler.

38. What is a "scaffolder" and how is it used? What are some of the challenges in scaffolding? What are some strategies? Give an example of a stand alone scaffolder.

39. What is the "Eulerian SuperPath Problem"?

40. In a "comparative assembly", each read must be mapped to a reference genome. What are the challenges in performing this mapping? Give examples of some techniques for this step, and discuss.

The following questions may require reading more than the assigned papers (or may not actually be fully answered in the current literature):

1. What is optical mapping, and how is it used?
2. Which type of assembler is most suited to short read technologies, and why?
3. Which type of assembler is most suited to the Sanger sequencing technology?
4. Describe some techniques for Assembly Validation.
5. Describe some attempts to do a "hybrid assembly", the different techniques that were used, and why they were used.
6. What is meant by "metagenome assembly", and how is it different from genome assembly? What special challenges does it present? What techniques are used to address metagenome assembly?
7. Consider two extremes in sequencing technologies - long reads with high error rates, and short reads with lower error rates. Which of these two extremes would be best suited for bacterial genome assembly, or human genome assembly? Why?
8. What is a chimeric read, and what causes it?
9. What is a chimeric assembly, and how is it created?
10. How does the "double stranded" nature of DNA complicate assembly?
11. In the context of genome assembly, what is a palindrome, and how does it complicate assembly?
12. What are some of the techniques to detect and correct errors in reads before assembly? How important is this?
13. Are errors in reads distributed uniformly throughout the read, or not? Discuss.
14. In a de Bruijn graph approach to assembly, every node is based on a (k-1)-mer, and edges are associated with k-mers present in reads. Discuss the issues in selecting the value for k, in terms of impact on accuracy, contig length, and running time.
15. What is the Lander-Waterman model, and what is it used for? Comment on its shortcomings.
16. Genomes vary in their nucleotide composition (some are GC-rich, for example). Does this impact assembly? Comment on this.
17. Some sequencing technologies are claiming to be able to create extremely long reads. Find out about this, and comment on it.