Instructions for the PhD Qualifying Exam of Luiz Irber

Objective (from Graduate Student Handbook)
The purpose of the qualifying examination is to assess a doctoral student's ability to perform several fundamental research related tasks including:

- reading and understanding relevant papers from the literature
- synthesizing ideas from separate papers into a coherent framework
- clearly expressing this framework in a written paper
- clearly delivering this framework in an oral presentation
- identifying possible extensions of the research described in the papers

In addition, the exam should serve as a chance for committee members to assess the background of the student (and possibly suggest remedies such as future coursework). The background component will be based both on mastery of topics related to the assigned papers as well as a mastery of a list of significant concepts presented to the student by the qualifying examination committee.

Instructions

For this exam, you are required to read and synthesize information from the following papers:

1. Daniel R. Zerbino and Ewan Birney. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* 2008, 18: 821-829. doi: 10.1101/gr.074492.107
2. Giorgio Gonnella and Stefan Kurtz. Readjoiner: a fast and memory efficient string graph-based sequence assembler. *BMC Bioinformatics* 2012, **13**:82 doi:10.1186/1471-2105-13-82
3. JT Simpson and R Durbin. Efficient de novo assembly of large genomes using compressed data structures. *Genome Res.* 2012 Mar;22(3):549-56. doi: 10.1101/gr.126953.111.
4. Heng Li. Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. *Bioinformatics* (2012) 28(14): 1838-844.doi:10.1093/bioinformatics/bts2

The length of your report should be between 4000-5000 words excluding references. You may structure the organization of your report as you like but its central theme should be about overlap graph-based sequence assembly in bioinformatics. Your report should include discussion of the following:

1. What is the basic procedure for overlap graph-based assembly (or string graph assembly)?

2. Compare the algorithms and data structures for building the overlap graph from a large number of reads.
3. What are the differences and disadvantages/advantages of the described approaches over de Bruijn graph based assembly?
4. How do the attached papers handle sequencing errors? How do they simplify the graphs by removing tips and bubbles? How do they take advantage of paired-end information?
5. Please describe one possible direction for future research.

You should be prepared to answer questions about the basic concepts in

1. De novo sequence assembly
2. Graph traversal
3. The Burrows–Wheeler Transform (BWT) and FM-index
4. Error correction methods during assembly
5. Design and Analysis of Algorithms