HiFiAssembler project in the multi-university *Bioinformatics* class (Winter 2021)

Independent research project. All students in the multi-university class will be able to participate in an optional research-focused "HiFiAssembler" project aimed at genome assembly using the recently emerged HiFi technology based on long and accurate reads. The HiFi technology, developed in 2019, is revolutionizing the way we assemble the genomes – it already contributed to the nearly COMPLETE human genome assembly by the Telomere-To-Telomere consortium in Fall 2020, the problem that remained unsolved since 2000 when the DRAFT human genome was first assembled by the Human Genome Project.

However, assembly of HiFi reads remains a poorly explored area with many open questions – there are still only a few HiFi assemblers and they all were developed very recently:

- HiCanu (Nurk et al., *Genome Research* 2020)
- Hifiasm (Cheng et al., *arxiv* 2020)
- Flye in the HiFi mode (Kolmogorov et al., *Nature Methods* 2020)
- LJA (Bankevich et al., *biorxiv* 2020)

Many issues in HiFi assembly, such as error correction of HiFi reads (Bankevich et al., 2020), remain poorly explored– can you develop a new HiFi assembler from scratch and address these issues? You will be provided with the sets of HiFi reads for E. coli, fruit fly, and the human X chromosome. Your goal is to assemble these reads, evaluate the quality of assembly, and prepare a short powerpoint presentation (maximum 10 pages) outlining your approach and summarizing your results.

The submitted assemblers are evaluated (based on the assembly quality, the running time, and the originality of the approach) and the winning assembly team in each class will make a short presentation for the entire class. Since various classes in this multi-university course end at different time, we will also identify the absolute winner (across all classes) when all classes finish in April 2021.

We encourage you to form multi-university teams working on the HiFiAssembler project (to promote interactions between students from different universities) but teams from a single university are also OK. However, we suggest that each team includes at most three students from the same university (but you can have as many 3*5=15 students in a multi-university team). Please select a name for your assembler that does not reveal your real names.

Below please find additional information about the project.

Datasets: You can find the original HiFi reads here:

- E. coli (95 514 reads, coverage 300x)
- Fruit fly (245 436 reads, coverage 50x)
- Human chromosome X (272 732 reads, coverage 32x)

When you debug your assembler, you will need to compare its results with the previously assembled sequences of these genomes that are available here:

- *E. coli* (4,641,652 base-pairs)
- Fruit fly (112,177,642 base-pairs)
- Human chromosome X (154,259,625 base-pairs)

We recommend that you use only the relatively small *E. coli* dataset for the initial development of your assembler and move to more complex datasets after you are satisfied with the results on the *E. coli* dataset. You may decide to randomly down-sample *E.coli* data to lower coverage (for example, to 50x) to minimize the memory footprint and running time - such down-sampling is unlikely to negatively affect the quality of your assembly.

To avoid possible over-training, the instructors will also test each submitted assembler on the fourth (secret) bacterial dataset that will be revealed only after the end of the project.

Start benchmarking from error-free read-sets! The error-rate in the HiFi reads is small (0.3% per a position) and as many 35% of all HiFi reads are error-free. However, these errors result in many *bulges* (see Chapter 4 of the textbook) and make the assembly task difficult. That is why, when you debug your assembler, you may want to start from error-free versions of the original reads. To generate these datasets of error-free reads you may want to align them against the reference genomes using minimap2 (Li, *Bioinformatics* 2018) or winnowmap2 (Jain, et al., *biorxiv*, 2020) and to identify the segment of the reference genome each read originated from. Afterward, you can substitute each HiFi read by this segment, resulting in an error-free read-set. For your convenience, we provide the set of error-free reads for human chromosome X.

We recommend that you move to analyzing real HiFi reads only after your assembler works well on error-free reads.

Evaluating the quality of assemblies. Please use the QUAST tool for evaluating the quality of assemblies to compare your assembly against the reference genome. Dr. Gurevich from Saint Petersburg University, the original QUAST developer (Gurevich et al., *Bioinformatics* 2013), will cover QUAST in his lecture “How do we compare LONG genomic sequences?” in this multi-university class.

Submitting the developed assemblers. For each interim submission, please run QUAST with each of the dataset you managed to assemble and send the QUAST report to mailto:abzikadze@ucsd.edu. Please use the unified minimal QUAST command ./quast.py -r reference.fa assembly.fa to generate the QUAST report. Sharing report.tsv for each generated report is sufficient. Each week, you can submit a single version of each assembly – please do it on Saturday. We will compile the leaderboard for each week. For the final submission at the end of the quarter, you will be requested to share the code of your developed assembler, allowing instructors to replicate your assemblies, create QUAST reports, and measure performance (CPU/memory).