Nanopore Sequencing Technology and Tools for Genome Assembly: Computational Analysis of the Current State, Bottlenecks, and Future Directions*

Damla Senol Cali[1,†], Jeremie Kim[1,3], Saugata Ghose[1], Can Alkan[2] and Onur Mutlu[3,1]

[1] Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA
[2] Department of Computer Engineering, Bilkent University, Bilkent, Ankara,Turkey
[3] Department of Computer Science, Systems Group, ETH Zürich, Zürich, Switzerland

† Presenting author (Email: dsenol@andrew.cmu.edu)

* The preprint version of the paper is available at https://arxiv.org/pdf/1711.08774.pdf

Nanopore sequencing technology has the potential to render other sequencing technologies obsolete with its ability to generate long reads and provide portability. However, high error rates of the technology pose a challenge while generating accurate genome assemblies. The tools used for nanopore sequence analysis are of critical importance as they should overcome the high error rates of the technology. Our goal in this work is to comprehensively analyze current publicly available tools for nanopore sequence analysis to understand their advantages, disadvantages, and performance bottlenecks. It is important to understand where the current tools do not perform well to develop better tools. To this end, we 1) analyze the multiple steps and the associated tools in the genome assembly pipeline using nanopore sequence data, 2) provide guidelines for determining the appropriate tools for each step, and 3) provide recommendations to use and develop nanopore sequence analysis tools. Based on our analyses, we make four key observations: 1) The choice of the tool for basecalling plays a critical role in overcoming the high error rates of nanopore sequencing technology. 2) Read-to-read overlap finding tools, GraphMap and Minimap, perform similarly in terms of accuracy. However, Minimap has a lower memory usage and it is faster than GraphMap. 3) There is a trade-off between accuracy and performance when deciding on the appropriate tool for the assembly step. The fast but less accurate assembler Miniasm can be used for quick initial assembly, and further polishing can be applied on top of it to increase the accuracy, which leads to faster overall assembly. 4) The state-of-the-art polishing tool, Racon, generates high-quality consensus sequences while providing a significant speedup over another polishing tool, Nanopolish. We analyze various combinations of different tools and expose the tradeoffs between accuracy, performance, memory usage and scalability. We conclude that our observations can guide researchers and practitioners in making conscious and effective choices for each step of the genome assembly pipeline using nanopore sequence data. Also, with the help of bottlenecks we have found, developers can improve the current tools or build new ones that are both accurate and fast, in order to overcome the high error rates of the nanopore sequencing technology.