High-Performance Computing for Assembly and Analysis of Big Genomics Data

René Warren, Benjamin VanderValk, Anthony Raymond, Shaun Jackman, Hamid Mohamadi, Daniel Paulino, Justin Chu, Ewan Gibb and Inanç Bilòr

British Columbia Cancer Agency, Genomics Sciences Centre, Vancouver, BC V5Z 4S6

Abstract

DNA Sequencing technology is developing at an unprecedented pace, surpassing the rate of advances in computer hardware development. Limited computer resources for storing, processing and analyzing vast data have spurred the improvements of file compression frameworks, lossy memory-efficient data structures, algorithms that use communication protocols for parallel programming, and adaptive approaches for handling large data on commodity hardware. Our research team oversees the development of such bioinformatics technologies. Past accomplishments include: creating the first assembly with millions of very short sequence reads (Warren et al. 2006), assembly of the human genome from short reads with the first parallel assembler (Gajewski et al. 2009) and last year, assembly of the then largest genome, that of the 20 Gbp white spruce (Brand et al. 2013). We discuss key enabling algorithms, specifically introducing data structures, processes, compression schemes within Abyss (Simpson et al. 2009), BBT (Chu et al. 2014), DIDA (Mohamadi et al submitted), Konnector (Vandervalk et al 2014) and TMOR (Warren et al. 2011) that are tailored to the needs of today’s big sequence data realm.

Funding

BC Cancer Research Foundation and Genome Canada. This research is also supported in part by Genome Canada, Genome British Columbia, National Cancer Institute of Canada and the U.S. National Cancer Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health under Award Number R01HG007182.

References

[Here the references would be listed in a bibliography format.]