Computational Challenges in Genomic Data Analyses

[1]Soumya Rao, [1]Timothy Cox, [2]Praveen Rao
[1]University of Missouri-Kansas City, [2]University of Missouri-Columbia

In 2003, a large international team of scientists made headlines by reporting the completion of the human genome project – a project taking some 13 years at an estimated cost of around $3 billion. Since then, rapid advances in sequencing technology and bioinformatics have made it possible for any small lab to sequence and analyze a whole human genome, or for that matter the genome of any species on earth, in mere days and at a cost of around $1000.

The breakthrough technology was "**short read**" sequencing developed by Illumina Inc. This approach, as its name suggests, is based on the random sheering of genomic DNA into short fragments (~300-500 base pairs) that are then sequenced from one or both ends to generate short, ~100-150 nucleotides read lengths. The millions of short reads are then bioinformatically assembled into large contigs. The short read lengths and the depth of coverage of the DNA by short reads contributes to the accuracy of the sequencing, that is, the ability to distinguish technical artifact and sequencing errors. Assembly of overlapping reads into contigs can be done *de novo* or, as more commonly achieved, by mapping to a reference genome if available. The accuracy of short read sequencing underpins its power in detecting single nucleotide variants and very small insertions/deletions that are commonly found to underlie disease phenotypes. However, a major limitation of short read sequencing is the difficulty in detecting larger structural genomic variants. Now, a third generation sequencing technology has emerged that generates **ultra-long reads** up to 100 kilobase pairs in length, thus being able to span and sequence complex and repetitive regions of the genome and detect large structural variants that may also be pathogenic. Variations on these technologies have also been developed to enable RNA sequencing and mapping of chromatin marks, in some cases even at the single cell level, providing insights into the spatio-temporal **expression and regulation of genes** on a genome-scale.

This genomics revolution, however, brought with it the challenges of handling, storing, transferring, and analyzing huge amounts of sequencing data: raw data generated from sequencing instruments can range from 50-100 GB/sample depending on the technology used and genome coverage required. The subsequent analyses of such large datasets impose additional burdens, typically doubling data storage needs and requiring significant computing power for processing (Fig. 1). For example, a *de novo* assembly from long read sequencing of a single 3 GB genome can take up to **3600 hr of CPU time** using **870 GB of RAM (**https://github.com/fenderglass/Flye**)**.

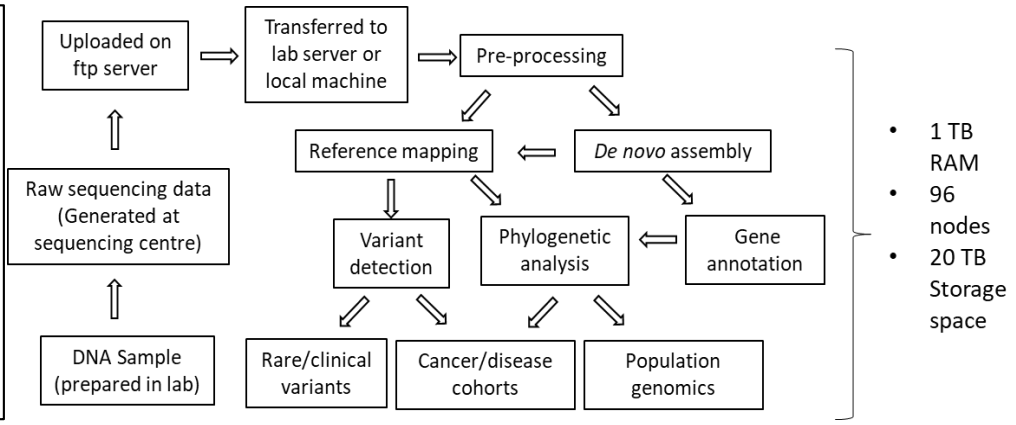

Fig. 1. The steps and computational requirements for a multi-sample genomics project using short- and long-read sequencing

The analysis of large-scale data like those in cancer studies, population-scale datasets, etc. have thousands of samples and involve comparisons within and between different groups in the search for the variants causing or contributing to a disease or phenotype. Moreover, the ever-expanding public databases like NCBI and SwissProt that have sequences and information on genomes, genes and proteins of all the organisms, databases like Pan Cancer Analysis of Whole Genomes Project, 1000 and 10000 genome projects for humans and other organisms, etc. have outpaced the ability of researchers to store, process and analyze the information contained in them. With the success of artificial intelligence (AI) and machine learning in data-driven decision making, new techniques and scalable software systems are needed to automatically gain insights from massive genomic datasets.

Some of the genome centers, national universities and research institutes have the necessary infrastructure and powerful computational resources and large, expensive **computing clusters** to manage and process, in reasonable time, the vast amount of genomic data generated every day. However, when it comes to local, regional and small universities and institutes, the cost and limitations of handling genomic data deter researchers from taking up a genomic study even if it is affordable and feasible in the present scenario.

For example, in our current set up, we need to keep switching between a regular desktop with 64 GB RAM/2 TB storage, which is quite slow but has sufficient space to store temporary files generated while running a genomics software, and a cloud experiment with 256 GB RAM/288 GB storage that finishes some jobs quickly but most do not finish due to the limitation of storage space. A substantial amount of storage space is dedicated to the basic bioinformatics and genomics software. The processing and analysis of genomics data from families with cleft lip and palate (CLP) is computationally intensive, challenging and inefficient with our current setup. Each raw short read sequencing file (~50-100 GB in size) needs up to 600 GB of temporary storage to enable mapping on the reference genome, creating a mapped read file of 150-300 GB. The subsequent steps for variant detection and comparison occupy up to 200 GB of storage. Thus, the genomic data generated for just a single family needs to be transferred to a local machine/server at each step so as to clear up the space for subsequent analysis, which has occupied 20 TB of server space so far. One short-term solution is to use a hyperconverged cluster infrastructure (HCI), which is the next-generation architecture for data centers. In an HCI, each node typically has a few hundred TBs of flash memory storage and tens of TBs of RAM along with gigabit networking. Thus, data movement can be drastically reduced compared to a traditional compute cluster leading to faster computations on massive genomic datasets.

There are some shared resources available like public, private or hybrid **servers and clouds** (Mashl et al., 2017) that may be used, but still these are beyond reach for most researchers due to budgetary constraints or lack of access, basic infrastructure, resources and knowledge of cloud computing, server administration, networking, security, and other development operations (Yakneen et al., 2020). There is a critical need for developing novel, accessible, efficient and cost-effective ways to improve computational capabilities and reduce the database size, processing time and storage footprint of genomic data in order to make sequencing a regular part of science and medicine. New techniques and genomic software are required to enable accessibility, efficient compression, rapid sorting and multi-dimensional indexing of genomic data and to eliminate the need for decompressing and downloading datasets from public or private servers or databases. In clinical setup, **precision medicine** can be achieved only if there is the feasibility of generating, storing and processing the huge amount of data from each and every patient using an analysis system that can be accessed through a regular desktop (Stein et al., 2015). The use of cloud–based computing clusters and software could be a solution for big data analysis, if it is made more accessible and economical for all. A large scale integrated grid project, pooling computational resources across the globe or even countries and states, may have great potential for addressing these issues and would be a huge step towards overcoming the major hurdles in genomic studies for individual researchers, regional universities and institutes. Therefore, we suggest NSF and NIH to invest in experimental testbeds for enabling academic research in genomics and precision medicine as these areas are challenged by massive datasets.

References:

Mashl, R. J. et al. GenomeVIP: a cloud platform for genomic variant discovery and interpretation. *Genome Res.* **27**, 1450–1459 (2017).

Stein, L. D., Knoppers, B. M., Campbell, P., Getz, G. & Korbel, J. O. Data analysis: create a cloud commons. *Nature* **523**, 149–151 (2015).

Yakneen, S., Waszak, S.M., Yakneen, S. et al. Butler enables rapid cloud-based analysis of thousands of human genomes. *Nat Biotechnol* (2020). https://doi.org/10.1038/s41587-019-0360-3