Delivering on the Promise of Precision Medicine

Rong Chen, Anthony Costa, Robert Klein, Patricia Kovatch
Icahn School of Medicine at Mount Sinai

Response to RFI on Science Drivers Requiring Capable Exascale HPC

The Specific Scientific Challenge

The insight gained from the integration of multi-scale data from new genomics technologies, national databases (e.g. dbGAP, TCGA), advanced electronic medical records, imaging, and personal biometric devices (e.g. FitBits) will transform healthcare. Precision medicine requires the dynamic correlation of multi-dimensional data of a specific patient with extremely large BioBanks with multi-modal, high dimensional data on tens of millions of patients. Currently, the ability to perform such analyses is severely limited by the scale and performance of available computational and storage resources. There is an urgent need for computational, storage and database systems to serve massive amounts of geographically-distributed databases data to many researchers at the same time. Taking full advantage of this scientific opportunity will require I/O performance and storage infrastructure several orders of magnitude greater than what's available now.

The Potential Impact

While modern medicine is based on treating the average patient, we now recognize that the molecular and physiological details of disease in any one patient can vary greatly from the norm. Through the integration of genomic and biometric technologies with electronic medical records, it will be possible to develop sophisticated predictors of disease onset and response to treatment, enabling more precise medicine. A reduction of unnecessary tests and ineffective treatments will save significant amounts of healthcare dollars for patients as well as for local, state and federal governments. Leveraging these vast datasets to identify at-risk patients and optimize treatment strategies will reduce morbidity and mortality from myriad diseases.

The Specific Limitations of Existing HPC

Genomic data processing workflows are I/O-intensive, read and write an enormous number of tiny files and require an extremely large number of Input/Output Operations Per Second (IOPS). This is very different from traditional HPC workflows that write and read from a few, large sequential files. Although a few technologies and techniques exist today that improve the efficiency of these workflows (e.g. flash), it is unclear if they will be affordable or scale sufficiently to handle exascale-sized datasets [1]. In addition, the exascale-sized genomics data will need to be correlated with geographically-distributed national data sets and from medical records and personal biometric devices. There is no current HPC capability available to stream datasets of this size from disparate locations, let alone with any sort of performance nor is there robust infrastructure and community-adopted best-practice methodology for integrating even single modality data sets at this scale.

Computational Parameters Expected in 2025

As a typical study in this field, we imagine a study following 1,000,000 people across the country: this number was chosen to be the same order of magnitude as the Million Veteran's Study from the VA or President Obama's Precision Medicine Initiative. For each participant, there will be a single genome sequenced, five tissues sampled and analyzed with RNA-seq and three epigenomic assays every six months for five years, and the use of portable biometric devices such as FitBit to record five physiological measurements every minute. While this study is reasonable in scale and the data could easily be generated over the next decade, the storage space required is immense. For each participant:

- 1 Human Genome 130 GB (Estimate from recent whole genome sequencing we performed on Illumina X10 machines)
- 1 RNA-seq profile 5 GB x 5 tissues x 10 timepoints = 250 GB
- 3 Epigenomic assays 1 GB x 5 tissues x 10 timepoints x 3 = 150 GB

5 64-bit biometric measures 8 bytes x (60 minutes/hour x 24 hours/day x 365 days/year x 5 years) = 0.02 GB

This is 395 GB/person x 1,000,000 people = 377 petabytes of raw input data for this single study. Analysis of the data through community-driven genomic sequencing pipelines (e.g., alignment to references, refinement, and variant calling) is an iterative process that at each step requires significant (though non-exascale) compute, doubling of storage, and massive I/O throughput for communication and storage of intermediate files. As the data produced scales linearly with number of tissues and timepoints, the data storage required for such a study would scale similarly as the scope of the study increases. Together the total data footprint will grow to over an exabyte for a single study of this magnitude, especially as the computational methodology is refined and analyses are rerun. On a typical machine, there will be many studies conducted with similar computational and storage requirements, necessitating the need for truly exascale storage and I/O performance. Current HPC I/O is in the terabytes per seconds range for block data. At this rate, it will take a million seconds or 11.5 days to read an exabyte of data, without the random reads and writes typical of the iterative "best-practices" genomics pipeline. It should be noted that the sample study described here only touches the surface of the range of tissues and timepoints that can be sampled using epigenomic and RNA-seq studies.

Other Capabilities Needed

New applications with the capability to stream I/O from local storage and remote, national databases in real-time will be needed to avoid the latency from reading and writing from hard drives. Flash can assist but it is currently cost-prohibitive and would need to be of a size large enough to store locally copies of geographically dispersed databases. Community-driven, open best-practices data management, security, curation, and distributed access are key requirements to the successful implementation of large-scale longitudinal studies like the one proposed as an example above.

Potential developments and pitfalls

Significant efforts have been devoted in recent years towards alleviating some of the pain experienced by the heavy I/O load leveraged on HPC resources by genomics-style workflows. These include efforts in envisioning computational genomics as a big data problem well suited for Hadoop-style resources, which is still in its infancy and has significant hurdles to overcome for efficient compute through a distributed data-driven style. Another style alleviates the random I/O bottlenecks (though not the streaming and data footprint limitations) by implementing pipelines using in-memory only models. These too have not been widely deployed. The research, results-driven and fast-paced style of development in this area have precluded the successful implementation of these kinds of approaches. I/O-heavy on-disk communication has continued to dominate the genomics space, and this will likely continue as computational methodology is refined over the next decade or more.

References

[1] Kovatch, P., Costa, A., Giles, Z., Fluder, E., Cho, H., Mazurkova, S. (2015). *Big Omics Data Experience*. Supercomputing 2015. DOI: 10.1145/2807591.2807595.