

CATALYZING THE ADVANCEMENTS IN GENOMICS TO LOWER BARRIERS TO SUSTAINABLE INNOVATION

The Genomic bio revolution has had a significant impact on our lives and economies due to acceleration in the development of computing, artificial intelligence, and automation. It has affected all of us, from our health to agriculture and consumer goods, energy, and materials. Considering the advancements in science and newer technologies, about 45 percent of the burden from prevailing global diseases could be addressed, impacting up to 60 percent of physical inputs¹. Having said that, we are still a long way away from attaining the full potential of these advancements.

One of the greatest barriers to attaining sustainable genomic innovation is pedagogic science laboratory courses. These courses need to be replaced with practical and discovery-based research courses. Such intervention is essential because our understanding of genomic science constantly changes, which becomes an intellectual and financial challenge for up-to-date access. However, this barrier can be addressed through collaborative national priority research projects.

Another barrier is the lack of adequate strength in the number of computation biologists, super computation systems and trained bio-computing experts that can perfectly analyze tera bites of genomic sequence data within a stipulated time framework.

The way revolutionary computing and digital electronics applications have transformed society over the past four decades, genomic and computing evolution seems to do the same. Looking back four decades ago, did any one of us imagine having access to any portals like Skype, Facebook, Twitter, or even have access to desktop computers back then? Today, cheaper sequencing is a key to uncovering that knowledge and catalyzing innovation in medicine, agriculture, drug development, vaccine development, and beyond. We have seen this impact in the development of Covid vaccine completion in less than a year.

An outstanding example of such progression is in genomics and computation biology, where DNA sequencing on a semiconductor chip can be done at USD 1000 per genome, which has otherwise consumed \$3 Billion to sequence one whole genome as a part of the Human Genome Project started in 1990². The genomics industry marked a new milestone that is not just big news for biotech and medicine but also exciting for all Technomists.

The pace of advances in genome sequencing technology has exceeded Moore's Law, as has been indeed said by Jonathan Rothberg, inventor of the Ion Proton technology. The connection between genomics and computing by sequencing Gordon Moore's genome sequencing has far better than doubled every two years since 2003 till date. Until a few years ago, Moore's law managed to stay ahead of the genomic curve due to exponential growth in storage and processing capacity of computers outpacing the genomic sequencing data. Nonetheless, after discovering Next-generation sequencing (NGS), genomics data has outpaced Moore's Law by four times.

Today, significant challenges revolve around storing big data of genomic categorizations that are more expensive to store, process and analyze than to generate. This challenge is not exclusive to genomic science alone. It has also impacted other diverse sectors such as physical science, finance, retail, and the Internet of Things, with worldwide data of 40ZB from 0.8ZB (a trillion GB) in 2009³. We are witnessing that widespread access to DNA sequencing has made it possible to share the genomic details of rare diseases and cancer, helping researchers identify links between mutations and various therapeutic responses. But there is also a growing concern about more genomic data that can't be used unless science knows what it all means, like a big chunk in the intronic region of our genome. The processing and analysis of the data deluge have spurred the dependability on cloud or utility computing (also known as elastic computing), where users can hire infrastructure-as-a-service through a "pay as you go" basis, thereby avoiding large capital infrastructure and maintenance costs. Due to advances in virtualization, such customized hardware and computational power can now be provisioned instantaneously using user-friendly web interfaces.

The solution does not lie in cloud-based computing alone. Big data presents problems of deviation from the traditional structured data that can be represented as semi-structured data such as XML, or unstructured data including flat files that are not compliant with traditional database methods. Cloud computing cannot address the challenge of big data analytics where large scale processing is required, particularly when the scale of the data exceeds a single machine. This is where genomics high-performance computing (HPC) and AI help. Take Lenovo's GOAST solution for instance. The HPC architecture called 'Genomics Optimization and Scalability Tool' (GOAST), is known to help researchers analyze a whole human genome in as little as 18 minutes, and whole exomes in about

¹ McKinsey Global Institute, The Bio Revolution; Innovations transforming economies, societies and our lives, May 13, 2020

² Alex Cabral, Blog, April, 27, 2019

³ J Biomedical Informatics, 46(5), October 2013

30 seconds, which is 160x faster than standard environments. Such accelerated execution speeds mean researcher get to process more genomes at higher throughput, find answers faster, and make breakthroughs that save more lives. Artificial Intelligence helps researchers to make sense of the difference between one genome and the other from the deluge of genomics data. Furthermore, the explosion of big data generated by the such advanced and sophisticated NGS machines has generated a significant demand for 'data scientists', computer scientists and mathematicians with expertise in big data, analytical techniques, statistics, data mining and computer programming.

In the 21st Century, if Big Data is used effectively in the health sector only, it can save 300 billion dollars per annum, as per the McKinsey Global Institute survey⁴. Though the Genomic science is experiencing big data overload, its benefit to humanity of deciphering such big biological data sets using NGS technology makes it the ultimate use case in the coming era.

Author Information:

Prof Jayesh Sheth
Founder and Chairman
Foundation For Research in Genetics and Endocrinology (FRIGE)
Institute of Human Genetics
FRIGE House
Ahmedabad, Gujarat
India
www.geneticcentre.org

⁴ Basel Kayyali, David Knott and Steve Van Kuiken. The big Data Revolution in US Health Care: Accelerating value and innovation, April 1,2013