IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

GIANT GENOMIC STORING SERVICE

Mr.Ashish L, Divya M V
Assistant Professor, MCA Scholar
Department of MCA

ABSTRACT : This is a project run under Google X. Google will store the genome in the cloud for $25, and the storage system could have a major impact on the scientific community. The hope is to collect millions of genomes to aid in scientific research. As MIT Review reports, the system could aid in collecting "cancer genome clouds" that would allow scientists to share information and run virtual experiments. Significant advancements in genomics analysis have been made possible with the emergence of better and more cost-effective tools. For example, cloud-based technologies, such as those offered by Google Cloud Platform (GCP), provide computational resources capable of analyzing massive amounts of genomic information at unprecedented speeds and in many cases, at a lower cost compared with on-premises solutions. Today, the use of cloud-based tools enables analysis across thousands of genomes to identify patterns and markers for disease predisposition, prediction, and causality. This helps improve how healthcare providers understand and treat disease,and creates better-informed treatment plans for patients.

Keywords : google cloud ,cloud storing, genomics ,human genome.

1. INTRODUCTION

The new initiative, called Google Genomics, is appealing to researchers in an effort to keep human genetic data in a secure cloud for $ 25 a year. MIT Technology Review recently reported that Google, in fact, had been rolling out the service silently for months, but Google's opposition to other health projects covered the news. The reason why Google Genomics can be so large is that a large amount of even personal

data comes from being able to compare larger genetic sets.

Google wants to create one major genomic site, such as, according to Deniz Kural, CEO of genome sequencing company Seven Bridges who spoke to MIT Tech Review, "if I could find lung cancer in the future, doctors would follow suit. my genome and my tumor's genome, then ask about the site of another 50 million genomes. The result will be "Hey, here is the drug that will work best for you."

Google is not the only one working on this project: Amazon is also, and a handful of young players are listed behind both information. MIT Tech Review states that Google already has at least 3,500 genomes from public projects on its servers and may have more in private projects.

2. LITERATURE SURVEY

The Literature survey offers the review of literature on Giant genomic storing in cloud storage by understanding basic concepts like human genome, genomic data science ,genomic sequencing, cloud storage

Shaila singh is a scientist and incharge of bioinformatics and high performance computing facility in Pune National center of cell science,discusses the application of machine learning in genomics. Machine Learning offers ample opportunities for Big Data to be assimilated and comprehended effectively using different frameworks. Stratification, diagnosis, classification and survival predictions encompass the different health care regimes representing unique challenges for data pre-processing, model training, refinement of the systems with clinical implications. The book discusses different models for in-depth analysis of different conditions. Machine Learning techniques have revolutionized genomic analysis.

David Wall and Catherine watcher discussed about performing Genome analysis using Amazon Web services and wrote a book on 2021 named “Genomics in AWS cloud” Get an introduction to Whole Genome Sequencing (WGS) , Make sense of WGS on AWS , Master AWS services for genome analysis Some key advantages of using AWS for genomic analysis is to help researchers utilize a wide choice of compute services that can process diverse datasets in analysis pipelines. Genomic sequencers that generate raw data files are located in labs on premises and AWS provides solutions to make it easy for customers to transfer these files to AWS reliably and securely. Storing Genomics and Medical (e.g., imaging) data at different stages requires enormous storage in a cost-effective manner. Amazon Simple Storage Service (Amazon S3), Amazon Glacier, and Amazon Elastics Block Store (Amazon EBS) provide the necessary solutions to securely store, manage, and scale genomic file storage.

“Sequencing technologies and genome sequencing is an article” of Chandrasheker pareek gives informations like The high-throughput - next generation sequencing (HT-NGS) technologies are currently the hottest topic in the field of human and animals genomics researches, which can produce over 100 times more data compared to the most sophisticated capillary sequencers based on the Sanger method. It gives idea about genomic sequencing.

Jitao yang published a journal of industrial information integration on Cloud computing for storing and analyzing petabytes of genomic data, Genomics next generation sequencing (NGS) and third generation sequencing (TGS) have a broad area of applications in life science, such as Non-Invasive Prenatal Testing (NIPT), ctDNA Testing for Non-Invasive Tumor Personalized Therapy, Whole Genome Sequencing (WGS), Whole Exome Sequencing (WES), RNA Sequencing, etc. The high throughput genome sequencing instrument is capable of sequencing thousands of samples in parallel, generating dozens of terabytes of genomic data in one day. The storage and analysis of the big petabytes of genomic data are approaching a very challenge for much of the biomedical research communities. In this paper, we give the design and implementation of a genomics cloud, which can scale storage and computing abilities flexibly and provides many easy to use genomics analysis software for customers. This paper gives the technical solution for building a genomics cloud based on CWL/WDL, Docker, DAG, NAS and Object Storage System. The implemented cloud platform also frees scientists from the burden of: building high performance cluster, managing millions of genomic data files, and scripting genomics analysis pipelines.

2.STORING GENOMIC DATABASE

GenomicsDB is the ultimate technology for a variety of genomic applications using advanced APIs provided in C ++, Java, and Python; One Multiple Data (SPMD) method. Specialists in both computer technology and genomics manage and store genomic data through various computer systems and software. Increasingly, data analysis and linking centers are part of research networks and provide these services. Genomic data generation requires significant financial support from institutions such as the National Human Genome Research Institute (NHGRI), which provides more than $ 125 million annually to support genomic data science efforts.

The generated data resources are often made available to the wider scientific community for easy data analysis. They organize and supply many kinds of information about the human genome, such as genetic and biological DNA. Many private and commercial cloud platforms work in partnership with government and civil society organizations, such as the National Institutes of Health (NIH), via the STRIDES program. These programs provide computer backup and management for genomic data and provide the necessary security and protection for the privacy of personal data in particular.

COLOMNAR SPARSE ARRAYS

GenomicsDB uses columnar sparse arrays where samples are polished into rows and genetic sites or different sites are mapped into columns. These columns are seamlessly segregated into thousands of devices, allowing shared genotyping in the Broad Institutes genome analyzer toolkit (GATK) to reach 100,000 or more samples. This allows bio experts to obtain analytical results with high statistical confidence. Low-level storage format enables faster and more efficient disk recovery compared to file usage.

SHARED-NOTHING ARCHITECTURE:

It is a distributed computing architecture in which each update request is satisfied by a single node (processor/memory/storage unit) in a computer cluster .The intention is to eliminate contention among

nodes.

4.EVOLUTION

The genetic field has improved dramatically since 1953, the year Watson and Cricks described the DNA structure of the double helix (Watson, 1953). Today, high-volume sequencing machines offer human genetic sequences within a few days, with high yields and low cost. As a result, millions of patients are growing worldwide each year. Their genome data are used to predict the type, environment, and progression of genes, rare diseases, diabetes, and cancer (Raheleh Rahbari, 2016) (Yuan Yuan, 2014). These developments have the potential to revolutionize health care as much as we can. In the near future, genetic data in conjunction with phenotype data will be quintessential in specific drugs, where targeted drug combinations will be integrated into each patient. Genome uses a combination of heuristics and the Smith-Waterman algorithm.

GENOME_WIDE ANALYSIS STUDIES (GWAS):

It is a way of scientist to identify inherited genetic variants associated with risk of disease or a particular trait.

GWAS that deal with a large set of samples face a commonset of challenges:

- Variant data is large and growing.
- Scalable and efficient retrievals are needed.
- Efficient transformations are needed.
- Using a scalable file system or ObjectStore.
- Creating a combined, indexed VCF/gVCF

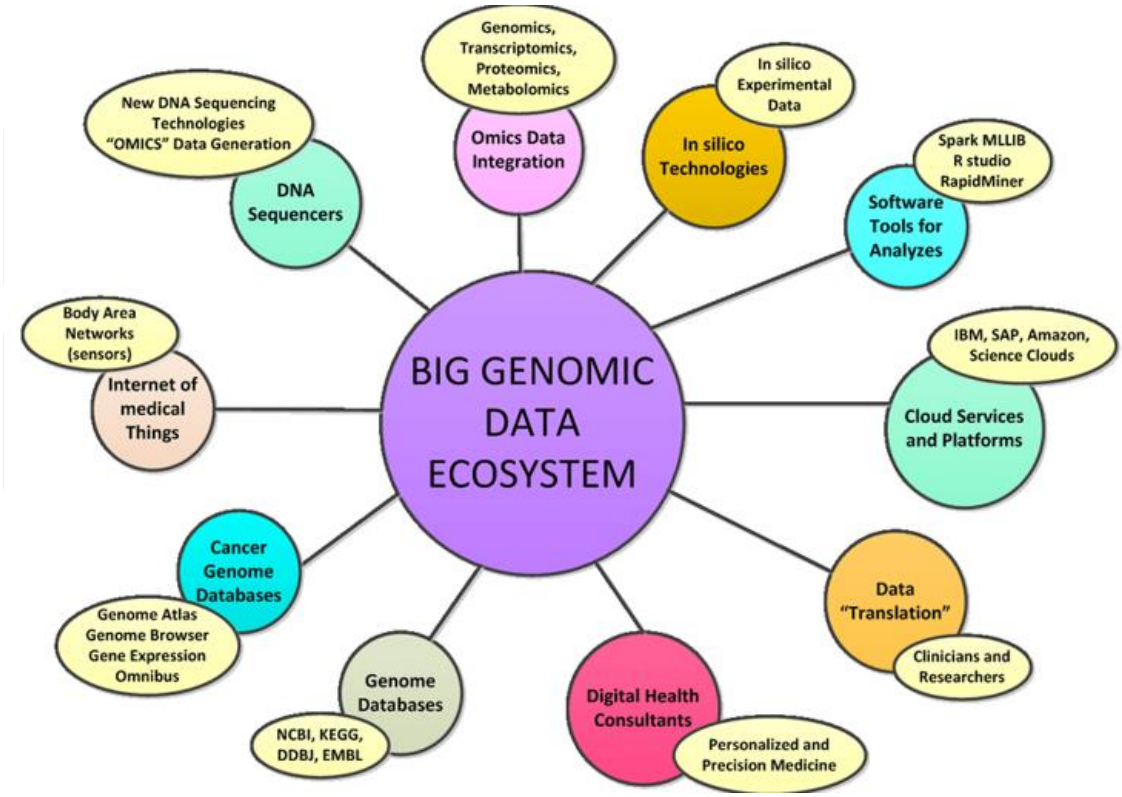
Genomic research is leading to deep new innovations in the healthcare industry. The new ideas make progress in risk assessment, diagnosis, prognosis, and treatment of patients. Genomics helps to make health care more personal by enabling targeted prevention and treatment programs that lead to improved clinical performance and outcomes. Organizations today use genomic research to develop more advanced therapies, provide diagnostic testing solutions, and modify traditional care delivery models.

5. GENOMIC DATA SCIENCE

Genomic data science is a field of study that enables researchers to use powerful computational and statistical methods to decode the functional information hidden in DNA sequence. Applied in the context of genomic medicine, these data science tools help researchers and clinicians uncover how differences in DNA affect human health and disease.

Genomic data science emerged as a field in the 1990s to bring together two laboratory activities:

- **Experimentation:** Generating genomic information from studying the genomes of living organisms.
- **Data analysis:** Using statistical and computational tools to analyze and visualize genomic data, which includes processing and storing data and using algorithms and software to make predictions based on available genomic data.



6.METHODOLOGY

SPARSE COLUMNAR ARRAYS

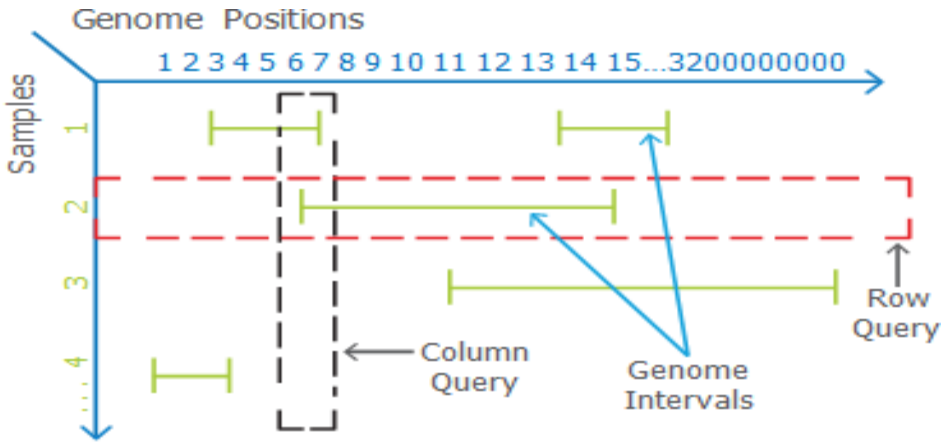


Figure 1. Variant data as a sparse 2D array.

The VCF file is organized by genome format; the shaded box in Figure 1 shows two VCF records. The first record shows that the insertion of the allele TA of chromosome 20 is found in the area 17960594. In the reference genome, the site contains only the nucleotide of T. The remaining record contains genotype, quality, and scores that may have been studied.The second record refers to the removal of 17986032, where the allele contains only nucleotide A instead of TA as found in the reference. Although the human genome has 3.2 billion characters long, mutations like these make up only 2 to 5 percent of its length. For this reason, separate data is less by nature; is stored in GenomicsDB using a two-dimensional (2D) data model for the same components.

Figure 1 incorporates a structured representation of the model. Here, the lines correspond to the person or sample and the columns correspond to the genetic makeup. Two examples of the above conversion will be stored in columns (17960594 - 1) or 17960594 and (17986032 - 1) or 17986031, respectively.

Extraction is required because VCF file locations are numbered from 1, but column references in GenomicsDB are numbered from 0. Each cell in the list contains multiple fields for storing genotypes, alternating allele, and metrics as quality and probability. points from an example of VCF records.

Tile db factors:

- Sparse variant data.
- Data can be stored as a 2D array.
- Uses a columnar mechanism.
- Efficient storage and retrieval.

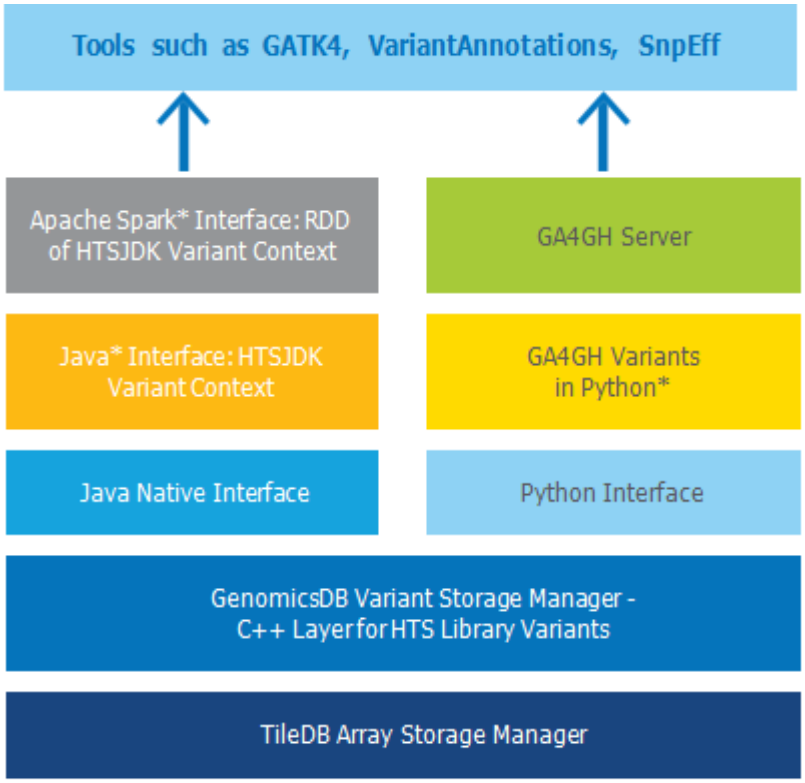


Figure 2. The different layers in GenomicsDB software stack.

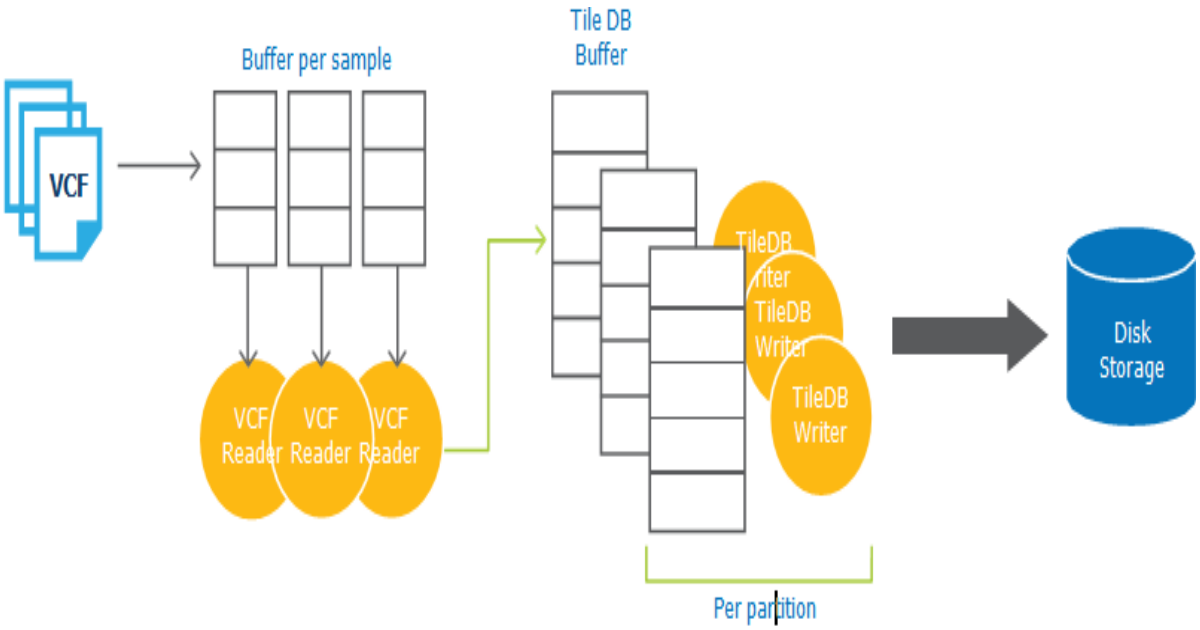


Figure 3. Data flow in GenomicsDB import.

ACCESSING GENOMIC DATA

7.BENEFITS OF CLOUD BASED GENOMIC DATA STORAGE:

- SECURELY STORE GENOMIC DATA

- As the use and opportunities for next-generation sequencing (NGS) continue to increase, so does the need for genomic data storage measurements. with speed and durability.Using this storage we can streamline our NGS data storage and analysis,increaseproductivity, and collaborate globally in cost effective manner.

- Save time and money while scaling production.

- Maintaining existing infrastructure with seamless integration.

- Access NGS data remotely to look at run trends or keep project moving.

- Spend more time on research and less time on infrastructure upkeep with automatedprocesses and user friendly options.

- Tap into massive processing power and scalability to analyse large NGS datasets.

- **It could help with the diagnosis and prevention of human disease.** Knowing more about the human body, we can understand how to treat and treat a variety of conditions. Even genetic conditions were incurable, and they were curable through the work of the Human Genome Project. Understanding our genetic profile means that doctors can diagnose conditions with the utmost certainty, or even rarely. This will lead to early adoption actions. Over time, the forecast level will increase.

- **It would allow us to modify medication for more effective treatment cycles.** • Medication before the genome was mapped was based on an all-inclusive solution. The drug worked for you, though it did not. One of the fastest growing areas in medicine is within the biopharmaceutical phase. Most of the advances that have come in the form of new medicines are a direct result of the work done by HGP over the past decade.

- **It could improve criminal justice proceedings.**

Our human genome is part of what makes each of us unique. Thanks to the DNA studies that were part of HGP, we have begun a process called “DNA fingerprinting.” By comparing DNA samples, one for a person and one collected, we have another way of identifying potential criminals. As the science behind this process continues to improve, our criminal justice systems can work better

.

List of the Cons of the Human Genome Project

- **It may cause a loss in human diversity.**

  What makes humanity such a strong race is its diversity. Although diversity can have negative components to it, such as genetic defects or mutations, it also strengthens us in numerous ways. Through diversity, we gain new perspectives. We have more creativity. We even have a stronger physical base for our overall genetic profile as a species. If we grow towards restricting the genetic pool for humanity instead of expanding it, then we may become weaker as a race.

- **Its information could be used to form new weapons.**

  Genetic information could be used to specifically tailor weapons to focus on population demographics. Once used, the weapons would focus on a certain genetic profile, eliminating all people with that profile from a civilization. That would reduce the amount of structural damage caused through conventional warfare, making it an attractive option for nations looking to secure more global resources.

- Inappropriate use of genomic data poses very specific risks since it can be used to identify an individual. Because each person can be identified by the variations in their genome, even databases of deidentified data can be used, in combination with other databases, to reidentify individuals.

- There is a strong argument for collecting and/or archiving genomic data in large databases. It is widely believed that “big” data will propel research forward. No single organization or laboratory will collect sets of data that will be large enough to truly accelerate the science that is critical to understand the genome.

9.Ethical, legal and societal implications of genomic data sharing

Conducting genomic research is consistent with a set of ethical obligations, as information about human genetics is associated with complex issues related to privacy and identity.

• Informed Permit: Researchers often ask permission from people for their consecutive genomes. But researchers should provide accurate information on how to use and share genome sequence data in an informed consent process.

• Confidentiality: Powerful computational tools can take sequential data from de-identified genomes and, under special circumstances, link them back to a person whose DNA sequenced. Investigators can use such

tools for useful purposes, such as identifying criminals who have left DNA behind in the crime scene. But the public benefits must outweigh the potential risks of using the data.

• Artificial Intelligence (AI): AI tools are increasingly helping researchers to process large amounts of gene sequences to look for hidden patterns in DNA. However, because AI algorithms tend to be less obvious, bias can go unnoticed when such algorithms are applied to DNA data. This area of genomic data science will require an in-depth study of ethics to navigate the unique differences between current methods of genomic data science (relying on human ingenuity to interpret results) and new AI methods. Although AI methods offer many promising benefits, and come to conclusions in completely different ways than humans, they therefore need to be carefully guided by ethical principles.

10. RESULT ANALYSIS

The purpose of our evaluation was to demonstrate our approach to improving the inclusion of genotyping inputs integrated with GenomicsDB and to demonstrate how implication of import and questioning measured by number of samples. We have selected 1,000 exome sequences for the 1,000 Genome Project. To perform scale tests, we divide the human genome into 16 parts. The method is to match the bytes labeled with each partition. The working time for submission or query is made on 1/16 of posts unless otherwise indicated. Therefore, different from the 1 / 16th human gene from all samples written or read in the test, the test set includes a set of four independent servers with a dual circuit,

TYPE	PARAMETER NAME	VALUES	DESCRIPTION
Import Configuration	num_parallel_vcf_files	1:N::1	Number of concurrent VCF files readers
	compress_tiledb_array	True or False	Enables compression
	segment_size	1:N::1KB	Buffer size to store TileDB cells in a columnar fashion; this buffer is used to both compress and serialize bitsfrom memory to storage via TileDB library
	size_per_column_partition	1:N::1B	Buffer size to store VCF records
	num_cells_per_tile	1:N::1KB	Number of array cells per tile in TileDB
	Partitions	1:1::1	Number of partitions based on genome positions or TileDB columns
Export Configuration	segment_size	1:N::1KB	Buffer size to read TileDB cells from storage to memory

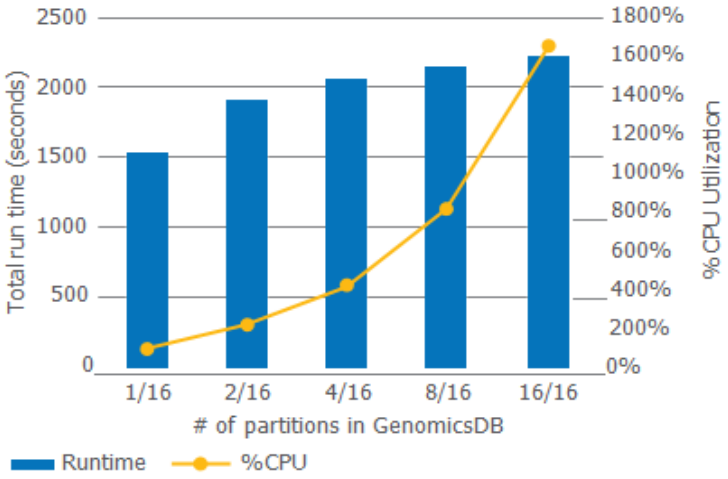


Figure 10. Read time with increasing number of partitions.

11.CONCLUSION

New inventions in genomic research open up new possibilities for understanding human health and for how diseases are treated. This section discusses current trends in genomic research and provides an overview of the remarkable laws governing the protection of sensitive genomic data in a clinical or research setting. Genomic data may contain personal, and possibly potentially sensitive, information about a person's physical characteristics and health. Genetic markers may indicate predictors of certain diseases or may be symptoms in certain cases. Organizations should adhere to the privacy and security practices of the genomic data in order to maintain the privacy of the individual.

12.REFERENCE

1. Adler, J.-l. G. (2013). Intel Genome Kernel Library. Retrieved from OTC Zlib Compression Library: https://github.com/Intel-HLS/ GKL/tree/master/src/main/native/compression/otc_zlib.

2. Ashley, E. A. (2016). Towards precision medicine. Nat Rev Genet, 507-522.

3. CoreGenomics. (2016, 05). How many genomes can the world sequence per year on X Ten? Retrieved from core-genomics. blog-spot.com: http://core-genomics.blogspot.com/2016/05/how-many-genomes-can-world-sequence-per.html.

4. DePristo, M. A. (2011). A framework for variation discovery and genotyping using next- generation DNA sequencing data. Nature Genetics 43(5), 491–498. GA4GH. (2017).

5. GA4GH Reference Implementation. Retrieved from GA4GH-Server API: https://github.com/ga4gh/ga4gh-server.

6. Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics.

7. McKenna A, H. M. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 1297–1303.

8. Papadopoulos, S. a. (2016). The TileDB Array Data Storage Manager. Proc.

9. VLDB Endow., 349-360. Raheleh Rahbari, A. W. (2016). Timing, rates and spectra of human germline mutation. Nature, 126-133.

10. https://www.intel.in/content/dam/www/public/us/en/documents/white- papers/genomics-storing-genome-data-paper.pdf (6/5/2022)

11. https://www.illumina.com/informatics/ infrastructure-pipeline-setup/genomic- data-storage-security.html (6/5/2022)

12. https://www.mobihealthnews.com/38069/google-genomics-storing-genomes-in-the-cloud-for-longterm-big-data- play (8/5/2022)