GenoVault: A Cloud Based Genomics Repository

Sankalp Jain
C-DAC: Centre for Development of Advanced Computing
Amit Saxena
C-DAC: Centre for Development of Advanced Computing
Suprit Hesarur
C-DAC: Centre for Development of Advanced Computing
Kirti Bhadhadhara
C-DAC: Centre for Development of Advanced Computing
Neeraj Bharti
C-DAC: Centre for Development of Advanced Computing
Sunitha Manjari Kasibhatla
C-DAC: Centre for Development of Advanced Computing
Uddhavesh Sonavane
C-DAC: Centre for Development of Advanced Computing
Rajendra Joshi (rajendra@cdac.in)
C-DAC: Centre for Development of Advanced Computing https://orcid.org/0000-0003-1299-0091

Software article

Keywords: Cloud, OpenStack, Genomics repository

Posted Date: November 17th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-105137/v1

Version of Record: A version of this preprint was published at BioData Mining on July 29th, 2021. See the published version at https://doi.org/10.1186/s13040-021-00268-5.

RESEARCH

GenoVault: A Cloud based Genomics Repository

Sankalp Jain, Amit Saxena, Suprit Hesarur, Kirti Bhadhadhara, Neeraj Bharti, Sunitha Manjari Kasibhatla, Uddhavesh Sonavane and Rajendra Joshi*

Correspondence: rajendra@cdac.in
HPC-M&BA) Group, Centre for Development of Advanced Computing (C-DAC), 411017 Pune, MH, India
Full list of author information is available at the end of the article

Abstract

GenoVault is a cloud-based repository for handling Next Generation Sequencing (NGS) data. It is developed using OpenStack based private cloud with various services like keystone for authentication, cinder for block storage, neutron for networking and nova for managing compute instances for the Cloud. GenoVault uses object-based storage, which enables data to be stored as objects instead of files or blocks for faster retrieval from different distributed object nodes. Along with a web-based interface JavaFX-based desktop client has also been developed to meet the requirements of large file uploads (>5 GB) that is usually seen in NGS datasets. Users can store as many as one million files in their respective object based storage areas and the metadata provided by the user during file uploads is used for querying the database. GenoVault repository is designed taking into account future needs and hence can scale both vertically and horizontally without any need for modification in the design. Users have an option to make the data shareable to the public or restrict the access as private. Data security is ensured as every container is a separate entity in object-based storage architecture also supported by secured file transfer protocol during data upload and download. The data is uploaded by the user in their individual containers that include raw read files (fastq), processed alignment files (bam, sam, bed) and output of variation detection (vcf). GenoVault architecture allows verification of the data in terms of integrity and authentication before making it available to collaborators as per user permissions. GenoVault is useful for maintaining the organization wide NGS data generated by experiments in various labs which is not yet published and submitted to public repositories like NCBI. GenoVault also provides support to share NGS data among the collaborating institutions. GenoVault can thus manage vast volumes of NGS data on any OpenStack-based private cloud.

Keywords: Cloud; OpenStack; Genomics repository

Introduction

Next-generation sequencing platforms are producing enormous data with the introduction of technologies like Sequencers [1]. Nucleotide sequence data are being produced at exponential rates leading to production of terabytes of data [2]. There is a need to store and organise such enormous data in a manner that enables easy access to data in a secure environment for the research community. Several traditional solutions are available in the public domain which uses distributed DBMS over file servers. Although relational databases have been primarily used in databases for many years, there are other solutions, such as Object Oriented and NoSQL used for storing datasets [3]. Security, scalability and integrity are the primary factors considered for understanding the advantages and limitations of the file-based, column

oriented design of relational database storage architecture. NGS data especially in the healthcare domain, is growing with a tremendous pace and the generated data needs to be stored in data repositories in an efficient manner. Depending upon the complexity of the experiment, the size of raw data generated by NGS also increases and can reach up to terabytes for a single organism, and for multiple samples, the size can increase upto petabytes [4]. A private cloud based repository can help an organization to keep the data locally and also share the data among its collaborating institutes. In order to improve the ease of access to such datasets, we have developed a user-friendly platform GenoVault for the retrieval and storage of NGS data useful for the scientific community. GenoVault is a software suite which enables cloud-based genomic repository facilitating quick archival and retrieval associated with integrated analytical engine support. GenoVault utilizes the full advantage of Cloud Computing viz. distributed computing and commodity storage. Users can upload the genomics sequence data onto the Cloud using the Web or JavaFX interface along with metadata which will be stored in a distributed manner on the Cloud. This feature enables swift container based efficient retrieval of the data. This centralized repository could be of enormous importance for healthcare and help in personalized medicine research. GenoVault can be deployed on any OpenStack [5] based public or private cloud infrastructure. Object Storage solution for storing genomics data is based on swift container for archival and retrieval of genomics data.

NEED OF GENOMICS REPOSITORY

As the size of genome increases the size of raw data generated from NGS also increases rapidly which can be of terabytes in size for a single organism. Heterogeneity in the data formats used by researchers for storage makes it cumbersome to share data among collaborating institutes. The generated genomic data files along with the metadata is stored in the genomics repository. Thus a centralized uniform repository would be of enormous importance for the researchers working in the area of Genomics [6]. Collaborators belonging to different laboratories can reuse the deposited raw NGS data and employ different protocols to analyse the data. This would ensure optimal utilisation of the data generated. Researchers would benefit with an infrastructure that ensures maximum accessibility, stability and reliability to facilitate working with and sharing of research data. Biological researchers carrying out bench-work generate NGS data but are unable to analyze the same because either there is lack of bioinformatics expertise or unavailability of compute and storage infrastructure. With data sharing, the research data does not remain restricted to the lab where it is generated. Dealing with the genomics data requires not only managing large data volumes but also being able to deal with the many different data formats (fastq; bam; sam; bed; vcf), query types, and real-time requirements. Users need to locate, understand, analyze and visualize the data to be able to use it effectively, which in the present scenario has scope to improve as there is a lack of suitable techniques, tools, and training [7]. There is an urgent need to create a genomics data infrastructure to help users to store their data and process it quickly, easily and effectively extract knowledge. These infrastructure should be supported with data mirroring and disaster recovery sites and follow a multi-tier

approach to address data consistency and data redundancy issues. The physical infrastructure should be supplemented by the data access, ownership policies and security considerations. In the coming future, it is expected that the sequencing of numerous species like humans, plants, animals, microbes, will be carried out. This will lead to a tsunami of sequence data of all species, which must be secured and stored in an efficient manner [8]. The infrastructure required for creating such a repository should be of massive scale. In order to manage this vast volume of data, we need to build an advanced genomics data archival retrieval system.

STUDY OF EXISTING GENOMICS PLATFORMS

Many customized common data repositories are available to help researchers working in a collaborative manner and yield high-quality research. Institutes like The National Center for Biotechnology Information (NCBI) [9], European Bioinformatics Institute (EBI) [10], DNA Data Bank of Japan (DDBJ) [11] provide an open platform for data sharing worldwide. International Nucleotide Sequence Database Collaboration (INSDC)[12], GenBank (at NCBI) [13]; ENA (EMBL EBI) [14] and DDBJ (at NIG) [15] have been serving as nucleotide sequence repositories for researchers across the world. These repositories apart from including raw sequence data also provide access to alignments, assemblies, and functional annotation. GenBank, EMBL, and DDBJ nucleic acid sequence data banks have from their inception, used tables of sites and features to describe the roles and locations of higher-order sequence domains and elements within the genome of an organism. The DDBJ provides a nucleotide sequence archive database and accompanying database tools for sequence submission, entry retrieval, and annotation analysis. DDBJ is administered by the Center for Information Biology and DDBJ (CIB DDBJ) [11] of the National Institute of Genetics. The EMBL Nucleotide Sequence Database is maintained at the European Bioinformatics Institute (EBI) [16]. The National Center for Biotechnology Information (NCBI) [9] resource is the most used worldwide and contains a variety of databases. The GenBank sequence database is an annotated collection of all publicly available nucleotide sequences. GenBank and other repositories receive sequences produced in laboratories throughout the world from over millions of distinct organisms. GenBank continues to grow at an exponential rate, doubling every ten months [17]. The analysis of such large data can help researchers to improve and excel in areas like human health, livestock, agriculture, and the environment. Many public cloud providers also provide genomics based solutions like Google Genomics [18], AWS Genomics [19] and Microsoft Genomics [20]. Genomics data is very well suited for cloud-based storage and analytics as the files are in text format and only a small percentage of output is needed for downstream analysis. Many cloud based genomics solutions also available like DNAnexus [21], Seven Bridges [22], DNASTAR [23], CLC Bio [24] etc. which provides various aspects of cloud from Software-as-a-Service(SaaS) to Platform-as-a-Service(PaaS).

Most of the researchers either deposit their data to the above mentioned repositories or keep the data with themselves in local storage. As most of the research data is not submitted in international repositories due to formalities, procedural issues, unpublished research etc. so there are chances that after a certain duration most of the data and research are lost. Hence there is a strong need to develop a

local centralized repository system where the researcher and scientist can store and retrieve their NGS data with ease besides supporting multi-organization collaborative research. Also once the researchers are ready they can submit the published NGS data from local private cloud to the international genomics repositories like NCBI, EBI and DDBJ.

GENOVAULT PRIVATE CLOUD REPOSITORY

GenoVault is a private cloud-based data storage that can use existing OpenStack based cloud deployment [25]. GenoVault can be implemented in an organization centric or a peer to peer collaborative manner. As genomics data is growing with the advances in high-throughput sequencing technologies, the size of raw data generated also increases. In order to improve the ease of access to NGS datasets, we have developed GenoVault with a user friendly interface for the retrieval of data for the scientific community [26]. GenoVault is a private cloud-based central repository for storing and retrieval of genomic data generated by various research groups. The solution is delivered in the form of a software suite along with support for analytical engines. GenoVault is based on OpenStack [25] cloud. It exploits and utilizes the full advantage of cloud computing, distributed computing and object based storage. Users can upload the sequence data onto the cloud using the Web or JavaFX interface of GenoVault along with metadata which is stored in a distributed manner on the cloud. This feature enables efficient retrieval of the data. The web client has been developed using JSF and jCloud APIs as shown in Figure 1. OpenStack cloud, jCloud API, FDT libraries, Object storage and JSF 2.0 are used for web-interface. Standalone client enables the transfer of large data files using Fast Data Transfer (FDT) [27] as shown in Figure 2. The standalone client has been developed using JavaFX interface and jclouds APIs which enables the transfer of large files using Fast Data Transfer (FDT) [27].

GenoVault uses OpenStack and the data storage part is implemented using OpenStack Swift [28] based on a distributed Object Storage solution. OpenStack key components include compute, storage (Cinder and Swift), and networking [29]. Swift offers cloud storage software that can store and retrieve data. Swift container scales and optimized for durability, availability, concurrency across the entire data set which is used for storing unstructured data that can grow without any bounds. Users can upload files using either web-based or standalone client along with metadata [30]. Metadata contains information regarding the type of sequencing platform used for sequence generation, number of samples, source organism, etc. as shown in Figure 3.

Uploaded files are accessible to the user and visible in the public domain. Users can search and download data if it has a flag for public access using a web-based user friendly interface. Standalone desktop client is capable of transferring files of large sizes. Users have their own area for uploading downloading data. NGS data files are stored in the cloud objects. The objects are stored in a distributed manner across Swift nodes. Distributed storage enables efficient retrieval of the genomics data.

Development was carried out using OpenStack as back-end with various services like nova, cinder, neutron, and keystone for authorization and authentication of the

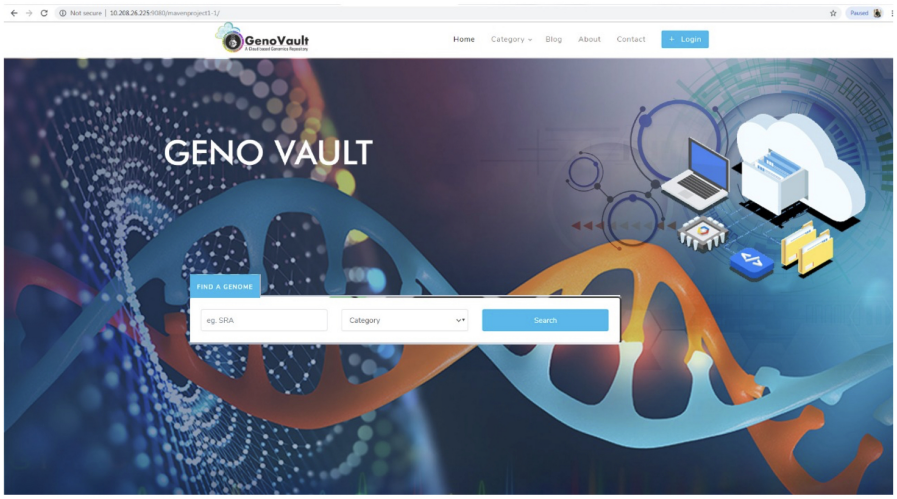

Figure 1 GenoVault Web Client

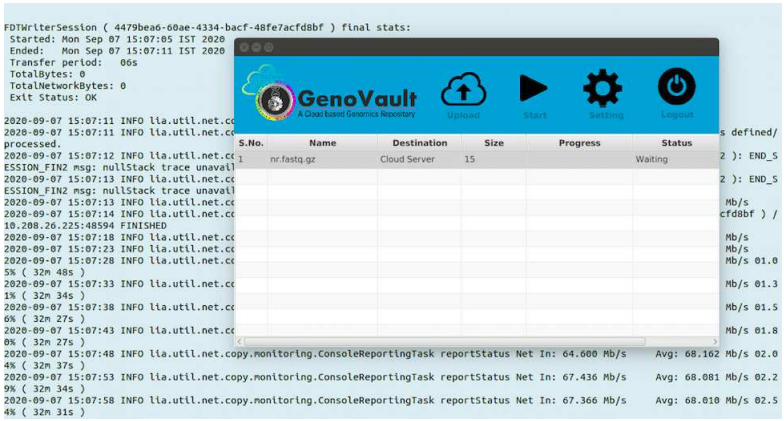

Figure 2 GenoVault file upload Client

user. Object based storage of data enables data to be stored as objects instead of files or blocks which enables faster retrieval from different distributed object nodes. These objects are stored into the container and each container is capable of storing one million files. GenoVault architecture allows verification of the data in terms of integrity and authentication before making it available in the public domain. GenoVault repository is designed taking into account future needs and hence can scale both vertically and horizontally without any need for modification in the design.

The facility is divided into components as shown in Figure 5 and described below.

Data Input

Data input involves gathering data from multiple heterogeneous sources like public genomics repositories and genomics data generated by various research labs. There are various formats of genomics data available as shown in Figure 4. Format conver-

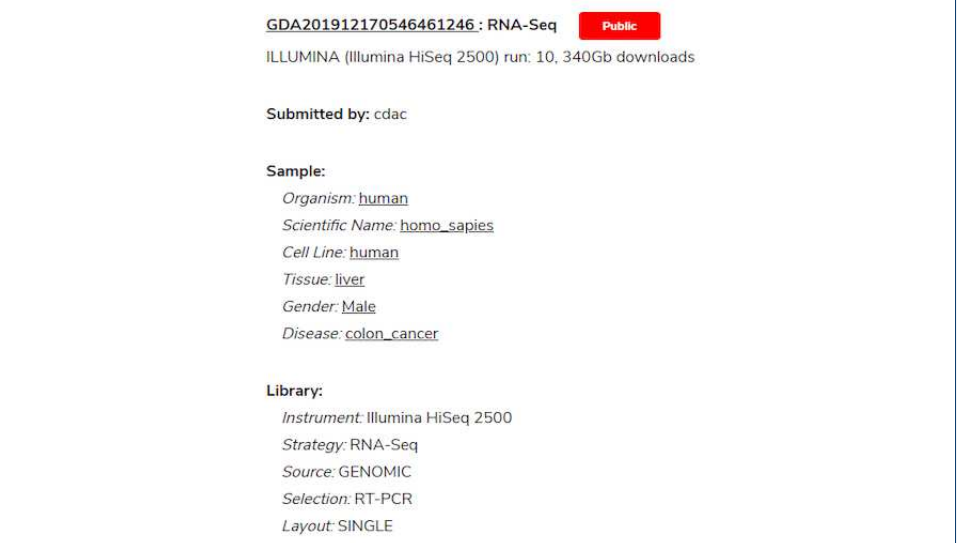

Figure 3 Metadata fields

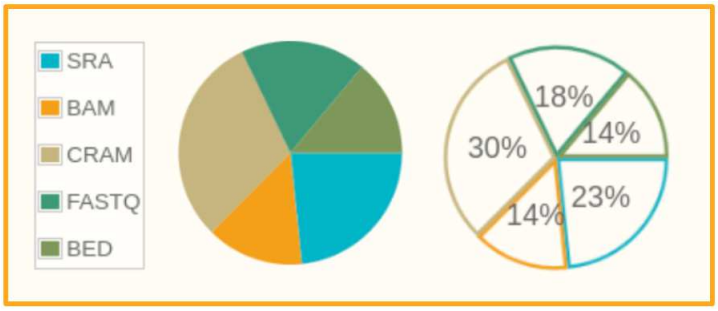

Figure 4 GenoVault supported file formats

sion at this step itself will make the retrieval operations like sorting, summarizing, consolidating, checking integrity, building indices and partitions easier. Data cleaning and data transformation are important steps in improving the quality of data and giving fast retrieval results. So data input is an important step facilitated by cleansing, translation, reforming and compression of data resulting in the validated data inflow for storage.

Access Layer

This layer accommodates all user related interaction applications, tools, hypervisor, firewall, software packages, software define network (SDN), virtual machines and volume storage [31]. This layer facilitates the user to set up all access controls secured by firewalls accessing SDN to volume/storage. It is directly connected to the primary storage with high speed networks like LAN and Infiniband.

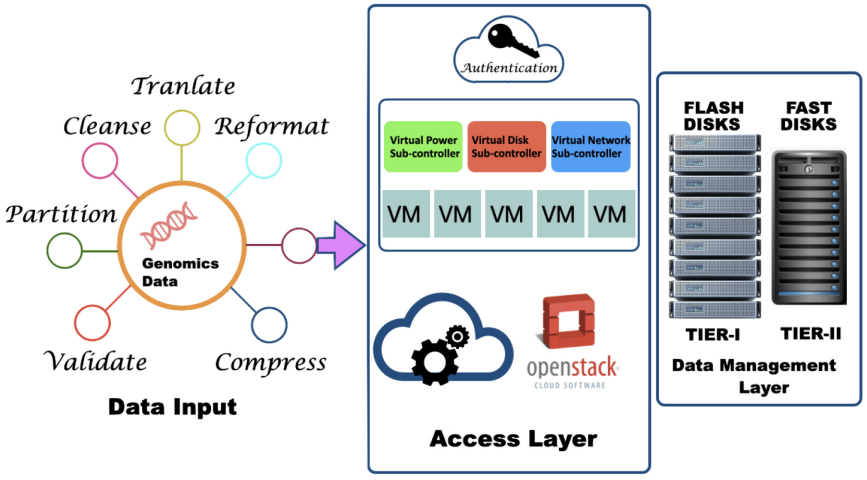

Figure 5 GenoVault Architecture

Data Management Layer

This layer has a multilayer mechanism which refers to configuring data storage infrastructure as a set of tiers, where each tier comprises a collection of media (memory, disk or tape) having distinctive performance, capacity and cost characteristics.

Primary Storage (Tier-I)

This layer facilitates the user to set up and seek all access control suggests that firewall and SDN to volume/storage. It is directly connected to the primary storage with high speed networks like LAN and Infiniband.

Secondary Storage (Tier-II)

It is the main and necessary element of the storage-as-a-service solution. It works as the main storage working under the principle of write once and read many times. In this, generally application and data are written once while reads are performed by several applications. It can offer information or data very quickly to be shared among all applications. It is well connected to the application layer through high speed networks like LAN and Infiniband and conjointly with storage devices.

Analytics

Extensive computational analysis using a number of algorithms and applications is required to infer scientific facts from Genomics Data coming through various research centres. We provide a platform for analysis of the Genomics Data using standard tools used in the respective domains. This enables researchers to get a preliminary overview of the trends in their data and hence serve as a good starting point for subsequent more extensive analysis. The analytics infrastructure consists of a Hadoop/Spark based node [32]. The proposed Hadoop node is deployed along with the cloud resources to exploit the benefits of Cloud technologies for the Genomics Data analysis. The Hadoop node denotes a fully functional hardware as well

as a software stack for Genomics Data analysis. Hadoop consists of the Hadoop Common which provides access to the file systems supported by Hadoop [32]. The Hadoop Common package contains the necessary JAR files and scripts needed to start Hadoop. The package also provides source code, documentation, and a contribution section which includes projects from the Hadoop Community. For effective scheduling of work, every Hadoop compatible file system should provide location awareness. Hadoop applications can use this information to run work on the node where the data is stored. The Hadoop Distributed File System (HDFS) [33] uses the information while replicating data to keep different copies of the data on different nodes for more fault tolerance.

SOFTWARE AND TOOLSET COMPONENTS

There are many technologies that are available to build software stack for ease of researcher's usability. Various technologies and platforms are used like Java, Cloud Computing (OpenStack [25], OpenNebula [34]), Object Storage Swift, Web Service, Swing, Struts, JSF etc. Every technology has some advantages and best technologies as per requirements are selected which fulfils the GenoVault's requirements.

Storing data

Big Data storage needs to be able to handle capacity and provide low latency for analytics work. It can be achieved by using hyper scale environments or NAS in a more traditional way. Very high-end enterprise cluster and SAN or Cloud environments with object storage are required for storing NGS data.

Moving data

Moving data between collaborators is also non-trivial and shipping hard drives is being used with poor internet bandwidth. A cloud-based gateway to scalable, high performance and open access analytics tools with close association with research centres to do a run-time analysis with genomics data and convey the desired results to the community. In cloud-based repositories data movement is minimized by availing the computation near data. It requires only input data and result data movement over the internet, all the intermediate data can remain in Cloud.

CASE STUDY AND SAMPLE DATA

The 1000 Genomes Project [35](1KGP) started in 2008 and completed in 2015, creating huge variation (with at least 0.01 of minimum allele frequency) and genotype data. This project was completed in four stages including the pilot phase. 1KGP includes 26 different populations, divided into five super populations namely African (AFR), Ad Mixed American (AMR), East Asian (EAS), European (EUR) and South Asian (SAS). The alignment files (in binary format) of 109 samples pertaining to Gujaratis in Houston (GIH) population were uploaded [36] into GenoVault as a case study sample data. Data size of 109 bam files is 850 GB, wherein the smallest file size is 7.7 GB and the largest file is 28 GB. These alignment files have been obtained after reference guided assembly using genome build GRCh38. The coverage of each of these samples is in the range 2-4 X. These files are suitable for variant discovery at cohort level. After uploading these files into GenoVault,

they were stored along with their corresponding metadata like accession details, sequencing platform, gender of the sample, population details. These metadata later aid in retrieval of the subsets as per user-requirements. As shown in Figure 2 we have observed a good performance in upload of data. An average speed of 68 Mb/s is observed while uploading the file through the FDT client.

Data Deluge

The challenge with NGS data is the large size of the NGS FASTQ files - the "data deluge" problem. The size of NGS data can be huge. For example, compressed FASTQ files from a 60x human whole genome sequencing can still require 200 GB. A small project with 10 - 20 whole genome sequencing (WGS) samples can generate 4TB of raw data [37]. Even these estimates do not include the disk space required for downstream analysis. Thus there is a need of managing the NGS storage with high throughput access mechanism.

DISCUSSION

GenoVault provides a complete infrastructure and ecosystem for the storage, management, retrieval and storage of Genomics Data. GenoVault has been developed by a multidisciplinary group of researchers from genomics as well as software engineers. GenoVault is useful in carrying out research pertaining to all aspects of genomics by providing solutions to problems which arise with increasing use and handling of genomics data. Development of a large cloud-based storage infrastructure dedicated for genomic data in conjunction with tools for advanced data archival retrieval of genomic data is the need of today. GenoVault is leading to creation of an advanced centralized genomic repository useful in rapid inference of the genomics data.

CONCLUSION

GenoVault is a cloud-based genomics repository in which the scientists, researchers and healthcare institutes can store and retrieve the genomics data for public or private access cloud-based infrastructure dedicated for genomic data. A common data repository using cloud technology will help researchers to work in a collaborative manner to yield high quality research. This will also shield biological researchers from the complexities associated with large data storage and provide associated access to high performance Big Data analytics node.

Availability of Software

The software is available to download at the link: (https://www.cdac.in/index.aspx?id=bio_products).

Acknowledgements

The authors acknowledge the Bioinformatics Resources and Application Facility (BRAF) at C-DAC, Pune. We also acknowledge Justas Balcas for providing permission to use the Fast Data Transfer (FDT) protocol libraries.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by Department of Biotechnology (DBT) research grant.

Ethics approval and consent to participate

Not applicable

Author information
Affiliations
Centre for Development of Advanced Computing (C-DAC)
Sankalp Jain, Amit Saxena, Suprit Hesarur, Kirti Bhadhadhara, Neeraj Bharti, Sunitha Manjari Kasibhatla, Uddhavesh Sonavane and Rajendra Joshi

Contributions
SJ developed the software code, AS designed the cloud framework architecture with manuscript writing, SH handled the backend cloud operations, KB did literature survey and planning, NB did the use case design and data availability, SMK did genomics data handling and manuscript writing, US and RJ were involved in the design, development and implementation phases. All authors read and approved the final manuscript.

Corresponding author
Correspondence to Rajendra Joshi.

References

1. Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P., Gu, Y.: A tale of three next generation sequencing platforms: comparison of ion torrent, pacific biosciences and illumina miseq sequencers. BMC genomics **13**(1), 1–13 (2012)
2. Gullapalli, R.R., Desai, K.V., Santana-Santos, L., Kant, J.A., Becich, M.J.: Next generation sequencing in clinical medicine: Challenges and lessons for pathology and biomedical informatics. Journal of pathology informatics **3** (2012)
3. Merelli, I., Pérez-Sánchez, H., Gesing, S., D'Agostino, D.: Managing, analysing, and integrating big data in medical bioinformatics: open problems and future perspectives. BioMed research international **2014** (2014)
4. Wandelt, S., Rheinländer, A., Bux, M., Thalheim, L., Haldemann, B., Leser, U.: Data management challenges in next generation sequencing. Datenbank-Spektrum **12**(3), 161–171 (2012)
5. Sefraoui, O., Aissaoui, M., Eleuldj, M.: Openstack: toward an open-source solution for cloud computing. International Journal of Computer Applications **55**(3), 38–42 (2012)
6. Tripathi, R., Sharma, P., Chakraborty, P., Varadwaj, P.K.: Next-generation sequencing revolution through big data analytics. Frontiers in life science **9**(2), 119–149 (2016)
7. Schadt, E.E., Linderman, M.D., Sorenson, J., Lee, L., Nolan, G.P.: Computational solutions to large-scale data management and analysis. Nature reviews genetics **11**(9), 647–657 (2010)
8. Buermans, H., Den Dunnen, J.: Next generation sequencing technology: advances and applications. Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease **1842**(10), 1932–1941 (2014)
9. National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – [cited 2017 Apr 06]. Available from: https://www.ncbi.nlm.nih.gov/
10. Madeira, F., Park, Y.M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A.R., Potter, S.C., Finn, R.D., *et al.*: The embl-ebi search and sequence analysis tools apis in 2019. Nucleic acids research **47**(W1), 636–641 (2019)
11. https://www.ddbj.nig.ac.jp
12. Cochrane, G., Karsch-Mizrachi, I., Takagi, T., Sequence Database Collaboration, I.N.: The international nucleotide sequence database collaboration. Nucleic acids research **44**(D1), 48–50 (2016)
13. Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W.: Genbank. Nucleic acids research **41**(D1), 36–42 (2012)
14. Li, W., Cowley, A., Uludag, M., Gur, T., McWilliam, H., Squizzato, S., Park, Y.M., Buso, N., Lopez, R.: The embl-ebi bioinformatics web and programmatic tools framework. Nucleic acids research **43**(W1), 580–584 (2015)
15. Tateno, Y., Imanishi, T., Miyazaki, S., Fukami-Kobayashi, K., Saitou, N., Sugawara, H., Gojobori, T.: Dna data bank of japan (ddbj) for genome scale research in life science. Nucleic acids research **30**(1), 27–30 (2002)
16. https://www.ebi.ac.uk/
17. Mizrachi, I.: Genbank: the nucleotide sequence database. The NCBI handbook [Internet], updated **22** (2007)
18. https://cloud.google.com/life-sciences
19. https://aws.amazon.com/health/genomics/
20. https://azure.microsoft.com/en-in/services/genomics/
21. https://www.dnanexus.com
22. https://www.sevenbridges.com
23. https://www.dnastar.com
24. https://secure.clcbio.com/myclc/login
25. https://www.openstack.org
26. Jimenez-Lopez, J.C., Gachomo, E.W., Sharma, S., Kotchoni, S.O.: Genome sequencing and next-generation sequence data analysis: A comprehensive compilation of bioinformatics tools and databases (2013)
27. https://github.com/fast-data-transfer/fdt
28. https://wiki.openstack.org/wiki/Swift
29. Khedher, O.: Mastering openstack (2015)
30. Bonthu, S., Srilakshmi, M., et al.: Building an object cloud storage service system using openstack swift. International Journal of Computer Applications **102**(10) (2014)
31. Jararweh, Y., Al-Ayyoub, M., Benkhelifa, E., Vouk, M., Rindos, A., *et al.*: Software defined cloud: Survey, system and evaluation. Future Generation Computer Systems **58**, 56–74 (2016)
32. Apache Software Foundation: Hadoop. https://hadoop.apache.org

33. Shvachko, K., Kuang, H., Radia, S., Chansler, R.: The hadoop distributed file system. In: 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), pp. 1–10 (2010). IEEE
34. https://opennebula.io
35. Siva, N.: 1000 Genomes project. Nature Publishing Group (2008)
36. https://www.internationalgenome.org/
37. https://stepik.org/lesson/189/step/1

Figures

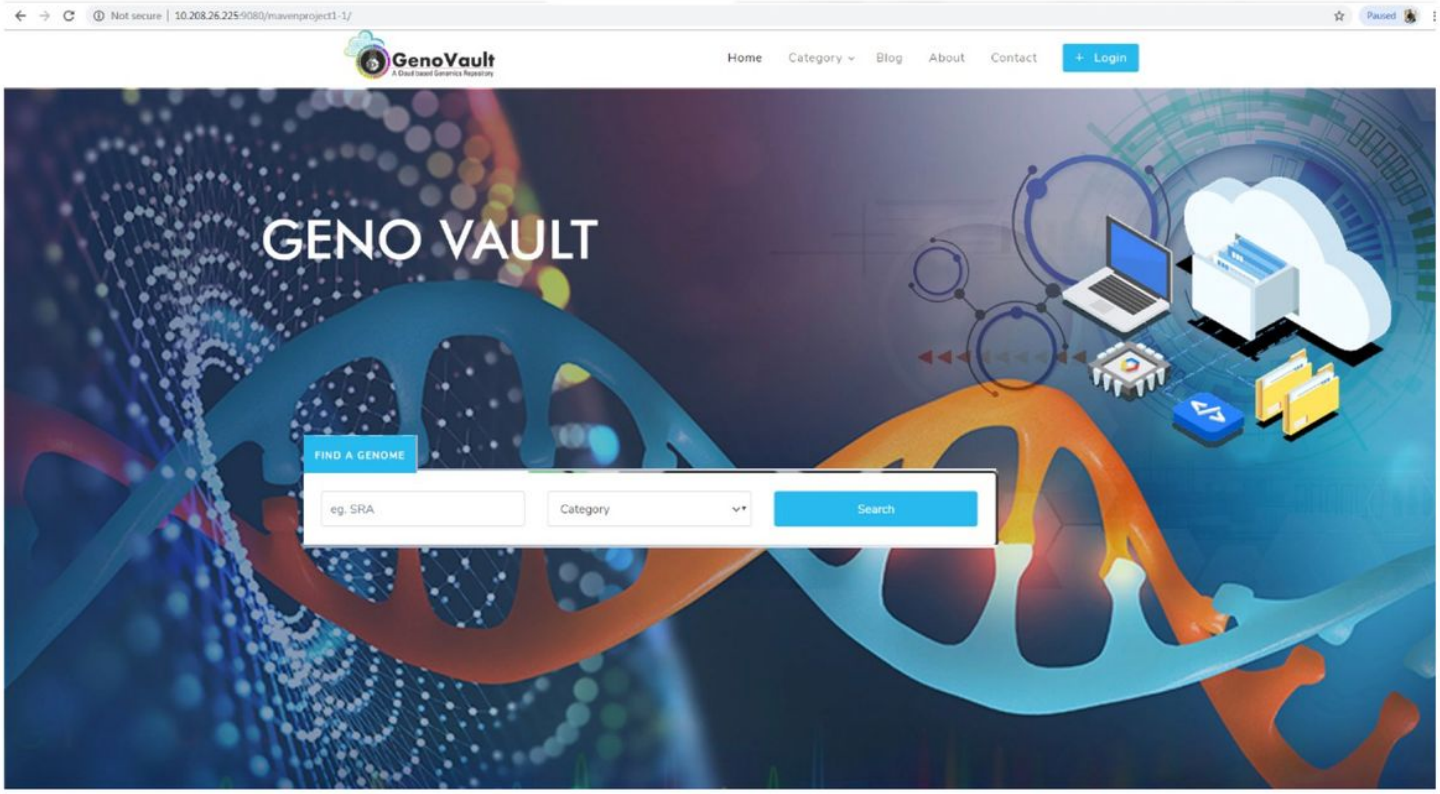

Figure 1

GenoVault Web Client

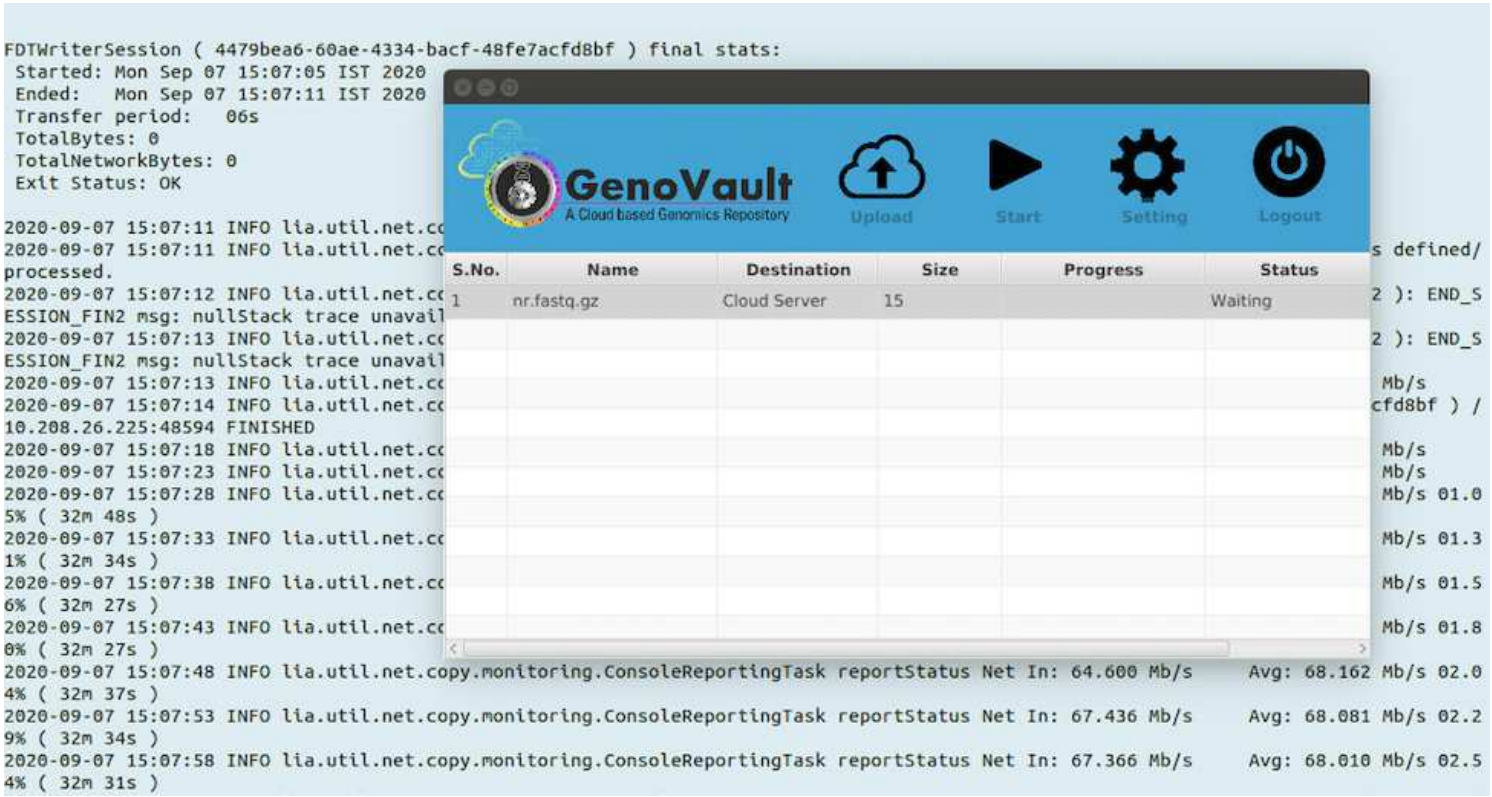

Figure 2

GenoVault file upload Client

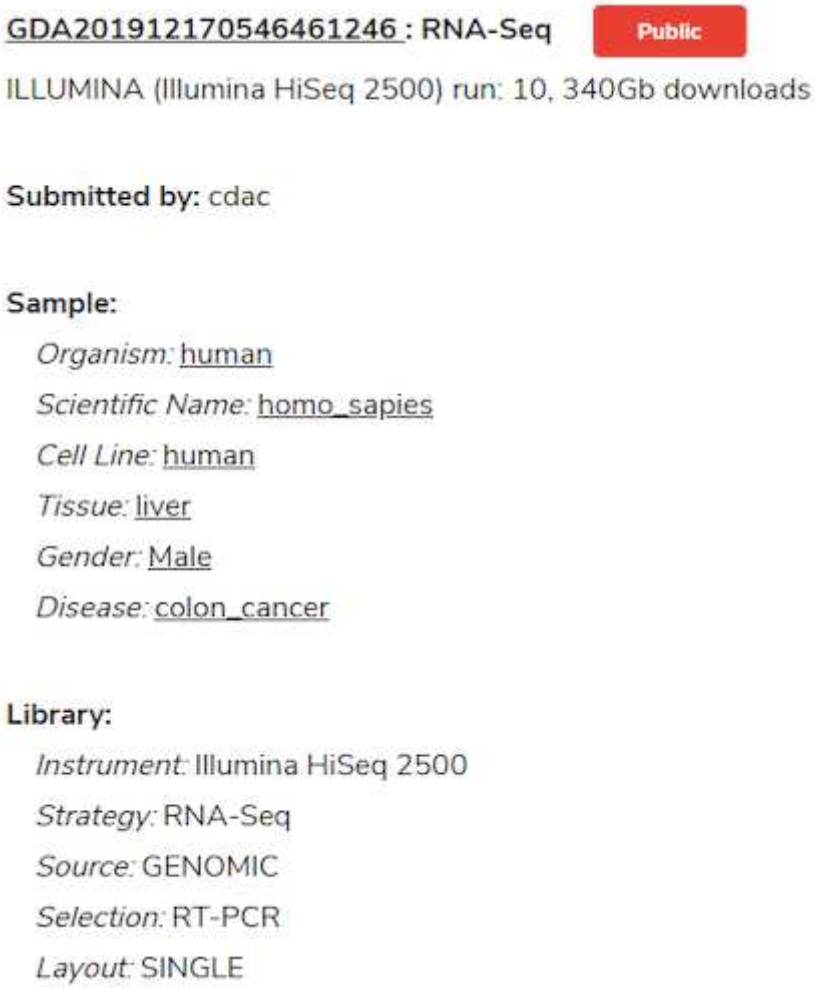

Figure 3

Metadata fields

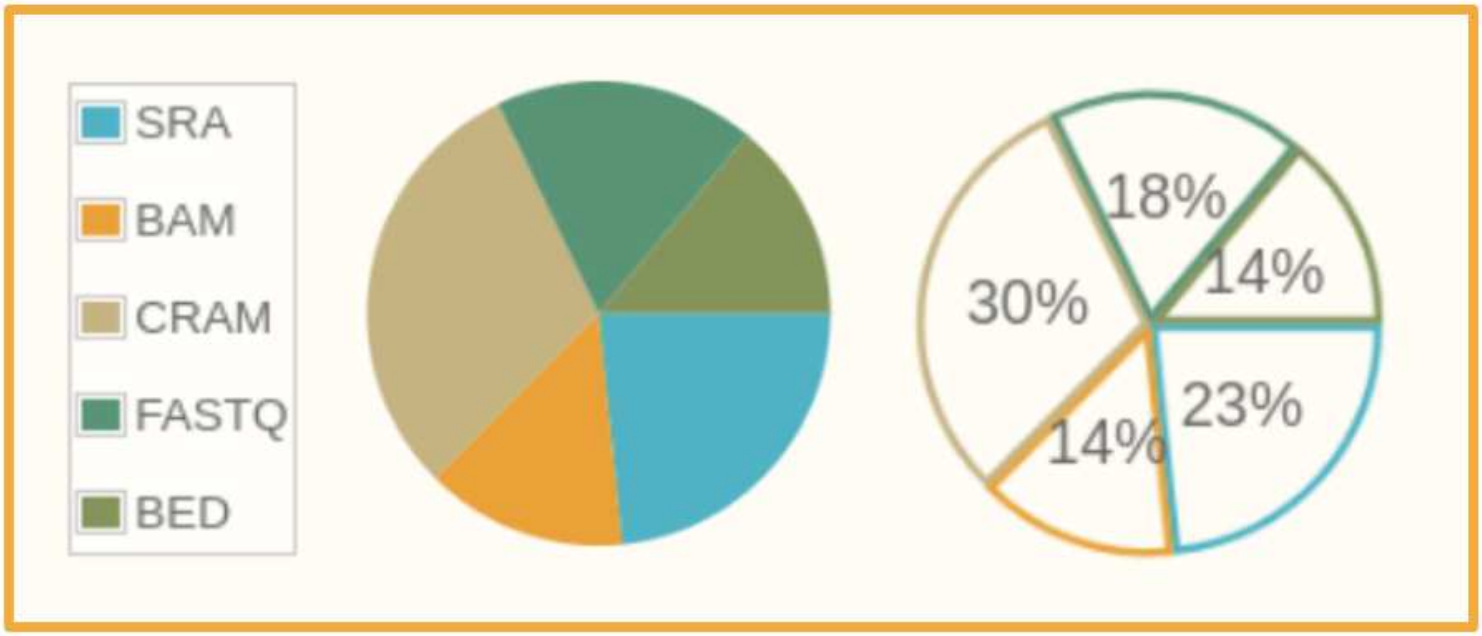

Figure 4

GenoVault supported file formats

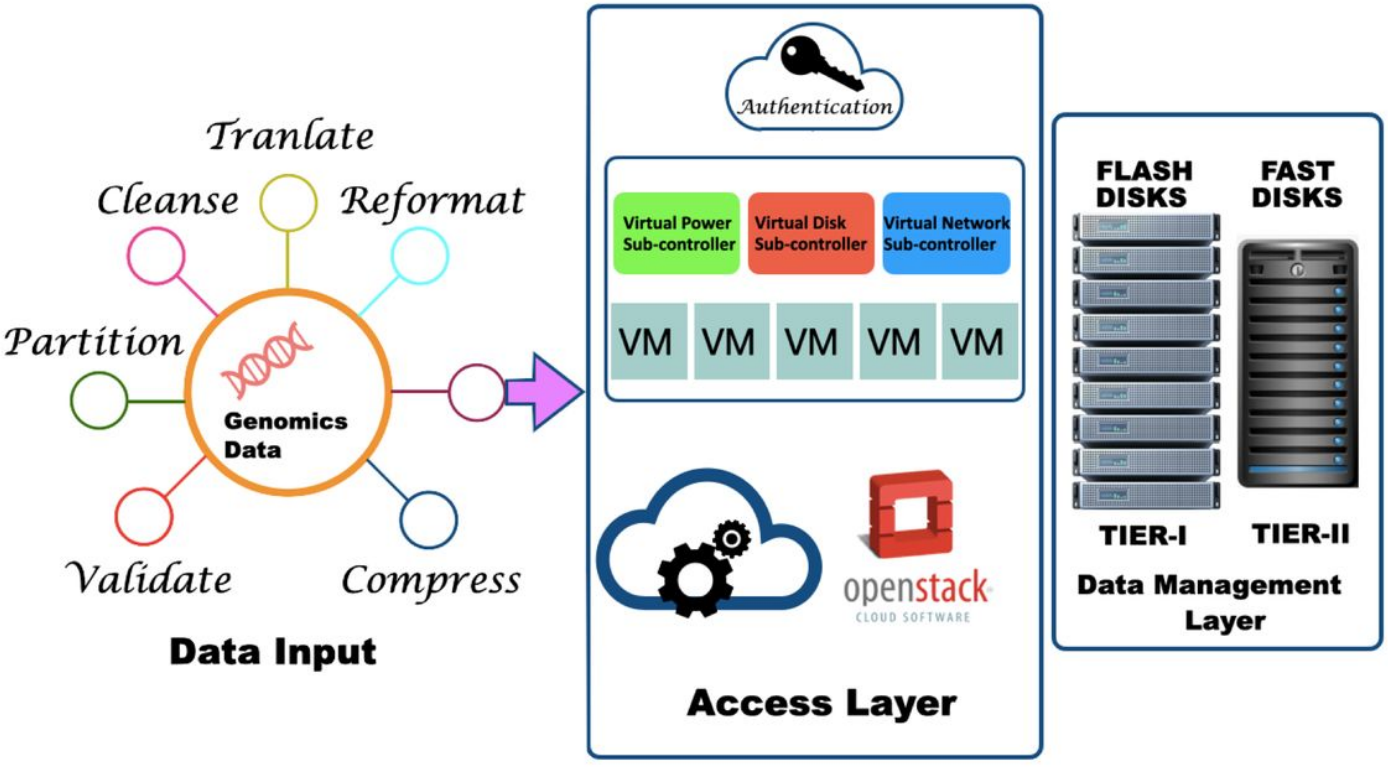

Figure 5

GenoVault Architecture