Forecasting the Future of Genomic Data Management

The past, present and future of genomic data

lifebit
lifebit.ai

Publication date: May 2023
Authors: Hannah Gaimster, PhD
Contributors: Ingrid Knarston, PhD and Amanda White

Introduction

Genomics is the study of the complete set of DNA in a person or other organism. DNA underpins a large proportion of an individual's health and disease status, therefore a genomic medicine approach is increasingly being applied in clinical settings.

Genomic medicine is where the study of clinical outcomes (measurable changes in health and well-being) is combined with genomics so researchers can better understand how a person's genome contributes to disease. Increasingly, advances in our understanding of the genome are contributing to improvements in disease diagnosis, drug discovery and targeted therapeutics.[1,2]

Worldwide, large population genomics sequencing projects are being established with the aim of implementing population-level genomic medicine.

These include programmes in the UK with Genomics England, the USA with the NIH's All of Us and Finland's FinnGen initiatives. With these projects comes huge amounts of genomic data which requires careful management to ensure this data remains secure yet accessible and analysable to researchers.[3]

In this whitepaper, we discuss how large scale genomic data is managed by researchers and organisations today and the challenges these current methods face. We consider new approaches and technologies that are being adopted to ensure genomic data management is effectively future-proofed going forward.

A data download or duplication approach is no longer sustainable

In the genomics field alone, there is now roughly 2 to 40 billion gigabytes of data generated each year, this makes data duplication or movement inefficient and expensive. Most importantly, data security and patient privacy are at risk every time data is moved.

Besides these inefficiencies, strict national regulatory frameworks (such as GDPR) that differ from country-to-country are making it near impossible to collaborate across borders using this traditional model of data sharing. Consequently, large-scale genomic data migration has become unfeasible.

"large-scale genomic data migration has become unfeasible"

Genomic data, then and now

How have researchers and organisations managed genomic data previously?

The first complete sequence of the human genome was described in the landmark Nature paper in 2001.[4] In the initial period of genomics research, relatively small genomic datasets were moved to a centralised location for analysis by researchers. Since then, sequencing technologies have continued to improve and become substantially less expensive.[5]

Additionally, there are now many more sequencing approaches available today - sequencing the transcriptome, proteome, lipidome, metabolome and more - we have truly moved from a genomics to multi-omics era, collecting more data than ever before.[6]

However, even today, many researchers accessing genomic data still begin their analyses by downloading this large-scale and highly sensitive data to a local or institutional computing cluster.[7]

This means organisations must maintain a set of computational tools to analyse the datasets they have downloaded. Further, it results in multiple copies of large-scale, highly sensitive patient data across many different sites.

Current approaches to genomic data management must adapt

There are several issues surrounding the approaches used to current manage genomic data. The main challenges are:

Maintaining adequate **security** and **compliance** around data privacy

Keeping pace with the ever **increasing amount of data t**hat is being collected by researchers

Ensuring that the data collected **can be combined** with other datasets for **scientific collaboration** to happen

lifebit
lifebit.ai

Security and compliance

Current approaches to managing genomic data typically involve DIY platforms or open source software solutions that can risk data privacy, as researchers must move the data to analyse it, making it vulnerable to interception. This is of particular concern for genomic data, which can be used to identify individuals.[8]

Going forward, organisations need to use advances in technology to develop safer, more secure platforms. Platforms should have the necessary infrastructure and controls to secure sensitive patient data, such as user authentication, monitoring, data encryption and pseudonymisation, role based access controls and ways to safely export data such as airlocks.

Compliance may also be compromised by the movement of data into centralised environments or 'data lakes' and across institutes/borders. Data transfer like this may not be fully compliant with laws like GDPR.

With advances like federated data-sharing technology, organisations can overcome these issues in a security-by-design model, where highly sensitive patient data can be made securely accessible by bringing researcher's analysis and computation to where the data resides, instead of moving data around.

Scalability

Advances in technology over the past two decades have led to a substantial decrease in the cost and time required for genomic sequencing.[9] It is estimated that over 60 million patients will have had their genome sequenced in a healthcare context by 2025.[10]

In genomics research, more data can lead directly to increased insights. For example, the first schizophrenia-associated genetic variant was identified using a cohort of 3000 individuals however a cohort 10x larger uncovered over 100x the variants. [11]

Furthermore, rare genetic conditions are, by definition, uncommon within the general population. Therefore, to maximise and validate insights from the data, researchers need to access multiple large-scale cohorts of clinico-genomic data to identify and validate findings. These cohorts are often in different places across the globe. In order to handle ever increasing amounts of distributed genomic data, we need to turn to elastic cloud storage solutions. [12]

To further compound these issues surrounding genomic data analysis, the data produced by many organisations around the world is often in different formats and not standardised to a common language. The data must be properly integrated and standardised, so it can be analysed seamlessly with other datasets. Most researchers struggle to combine their in-house data with external data sources to increase the power of their analyses, with 64% of health data users stating they do not have the expertise to easily standardise data.

The increasing size, complexity and disparate nature of today's datasets is projected to continue growing as the cost of genome sequencing continues to fall. This data-rich ecosystem we have today is no longer compatible with moving data to a centralised location for analysis.

Limited collaboration and usability

Traditional methods of accessing genomic data via copying or moving data actively limit collaboration. Researchers struggle to access data, typically having to wait many months for access.

Furthermore, data is commonly only accessible via inflexible DIY platforms, which are typically only accessible to data scientists and researchers well-versed in coding. In order to address this limited usability of the data and in-line with broader trends in the software industry, no-code solutions are growing in popularity. One example of this is DepMap that provides a straightforward graphical user interface to understand cancer vulnerabilities from available chemical and genetic perturbation data.

With point and click interfaces and intuitive workflows, platforms like theses are designed to make it easier for researchers with limited coding knowledge to access and analyse these data.[8]

When researchers are limited in the size and diversity of cohorts they can securely access, they are stifled in terms of collaboration and commercialisation opportunities. An important approach to improve collaboration is by using federation to securely link and access disparate datasets.

Overcoming the issues surrounding secure data access

Trusted research environments and federation provide a secure way to access genomic data

Pharmaceutical companies and national precision medicine initiatives are turning to innovative solutions that take analysis to where distributed data lies, avoiding costly and risky data movement across platforms and borders.

Many public sector organisations, government bodies and healthcare providers, such as the NHS in the UK, have employed trusted research environments (TREs) to ensure patient privacy and security across sensitive genomic data. TREs are secure computing environments that allow approved researchers from authorised organisations a safe way to access, store, and analyse sensitive data remotely.

TREs can be securely linked via federated technology, so researchers can easily access, collaborate, and analyse disparate datasets without data movement.[8] Federation of TREs can enable authorised researchers to securely combine distributed data to run joint analyses from anywhere in the world.

> "Federation of TREs can enable authorised researchers to securely combine distributed data to run joint analyses from anywhere in the world"

The Lifebit TRE enables organisations to overcome the three main challenges of security, scalability and collaboration in genomic data management.

Parker Moss, CCO of Genomics England

"

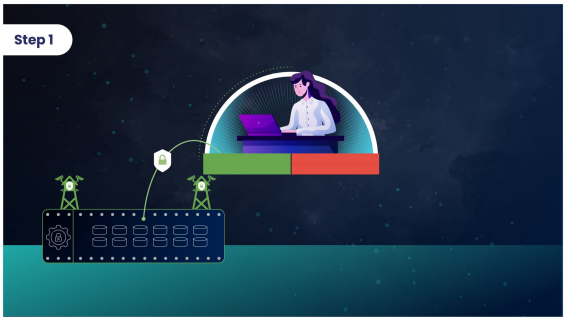
A challenge facing the precision medicine field lies in data governance. With strict national regulatory frameworks, there's a real pressing need to leave data at rest, but to **analyse it alongside international datasets** and integrate it through **federated links**.

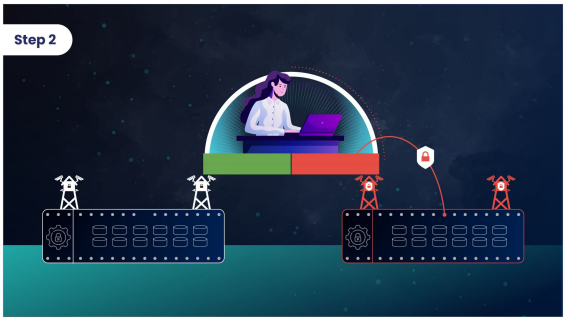
Lifebit's technology addresses this issue – **keeping patients' data secure and privacy** protected in our dataset, while enabling researchers from academia and pharma to **analyse this data collaboratively** in conjunction with their other complementary datasets. That's a very powerful value proposition.

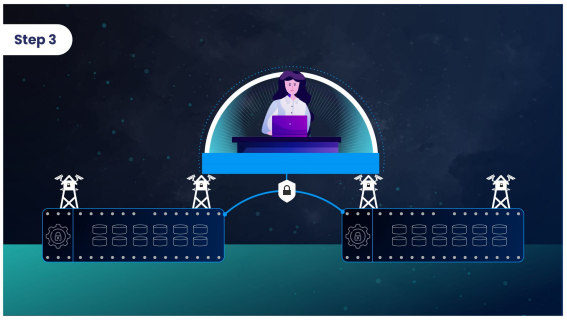
lifebit

Trusted research environments and federation provide a secure way to access genomic data

These graphics shows how federated data analysis works. Traditional methods of data access involve researchers downloading data to an institutional computing cluster (steps 1 and 2). With federated analysis, the analysis is brought to where the distributed data lies (step 3).







Critically, federation brings computation and analyses to where the data resides, thereby eliminating the risky movement of data and removing many existing barriers to accessibility.[13] Such technology means that data can be made securely accessible but that data stewards (eg biobanks and healthcare providers) retain jurisdictional autonomy over data, a key concern within the context of international data sharing.[9]

Recently, a pioneering example of multi-party federation between the TREs of Genomics England and National Institute for Health and Care Research (NIHR) Cambridge Biomedical Research Centre (BRC) was performed. This allowed secure data analysis across TREs in the UK's first known demonstration of genomic data federation. This highlights that the technology now exists to facilitate secure data access via federation for authorised researchers.

Data standardisation and cloud computing support increasing complexity and size of genomic data sets

International initiatives are beginning to come together to tackle the issue of limited data collaboration and interoperability. One example of this is The Global Alliance for Genomics and Health (GA4GH) which sets standards to promote the international sharing of genomic and health-related data, in part by setting interoperability standards and providing open-source APIs.[14]

Common Data Models (CDMs) are crucial to ensuring data is interoperable, with several growing in popularity in the life sciences sector recently including OMOP CDM from the OHDSI (specifically for clinical-genomic data) and Study Data Tabulation Model (SDTM) from CDISC.

Additionally, extraction, transformation, loading pipelines (ETL) pipelines that can automate this work to process and convert raw data to analysis-ready data help further simplify this process for researchers.

Normalising all data to internationally recognised standards allows researchers to perform joint analyses across distributed datasets, key to ensuring diversity and representation of as many populations as possible in studies.[8]

With the ability to process immense datasets, computational resources are an important consideration. The scale of distributed multi-omics and clinical datasets available today has brought an increasing shift towards commercial cloud infrastructure. The 'elastic' nature of cloud computing means researchers only pay for what they need. Furthermore, cloud computing builds capacity for state-of-the-art capabilities in encryption, firewalls and monitoring. [12]

No/low code tools will democratise access to genomic data worldwide

The software industry is currently shifting towards “no/low-code” tools to support a wider range of end users with and without a data science background, thus enabling full democratisation of access to genomic data and the insights derived.[8]

The Galaxy Community, an initiative within ELIXIR, is one such example offering a web-based platform to facilitate computational research for a variety of ‘omics’ types. [15] These types of tools enable users of diverse backgrounds to view the data directly or build reproducible pipelines and complex workflows for analyses.

While such low-/no-code tools are a huge first step, there should ideally be an end-to-end, federated solution for researchers as well as clinicians - providing the latter with the resources they require to understand their patients’ data.[16, 17]

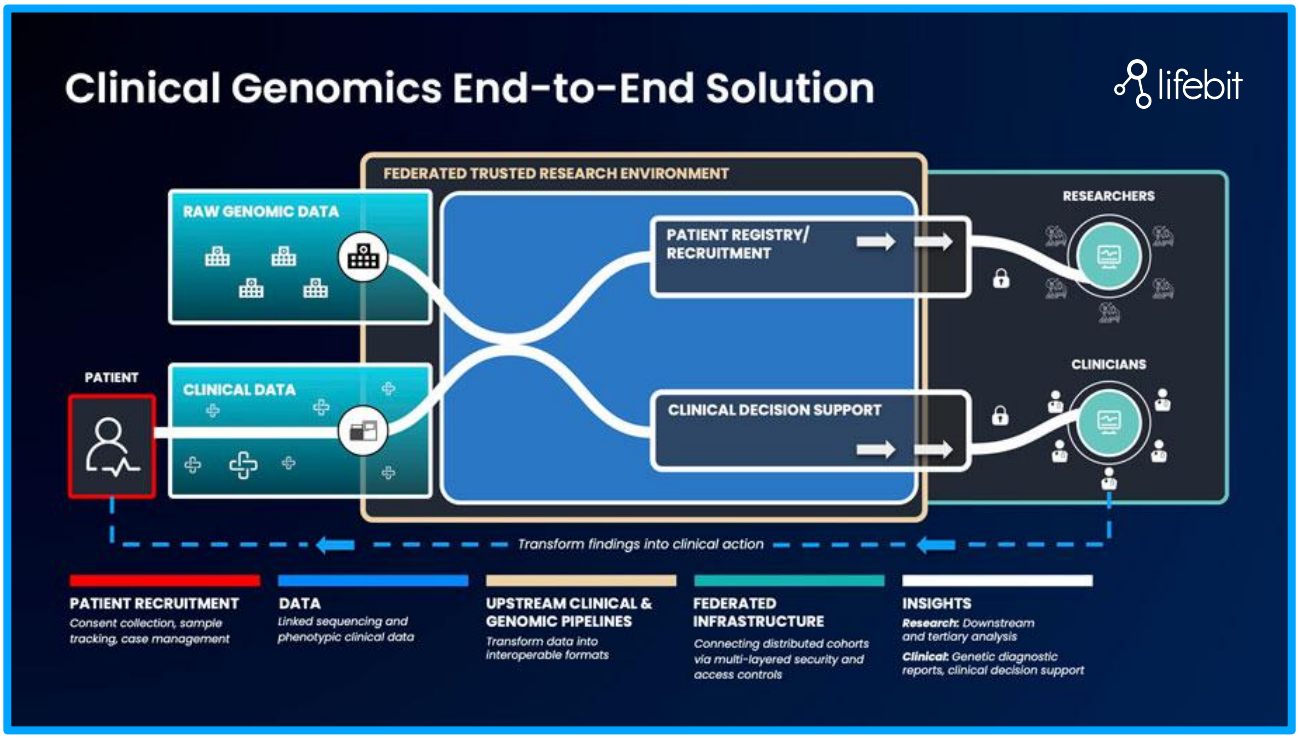
An end-to-end data platform, as shown below, can build upon the current advances of federated data architectures and be capable of ingesting clinical and raw genomic data, can help democratise access and accelerate the generation of clinically actionable insights.

No/low code tools will democratise access to genomic data worldwide

In an end-to-end solution, genomic data is collected and fully standardised into interoperable formats. Next, these data are ingested into the cloud-based federated architecture, to allow authorised users to access and combine this data with other disparate sources to build unique and valuable analysis cohorts.

Strict security measures will facilitate results export to authorised clinicians and researchers, to enable them progress therapeutic discovery and make informed clinical decisions.

It is possible to imagine that in the near future a platform like this could potentially securely integrate between a country's healthcare network, national genomic medicine initiatives and sequencing laboratories.



	Current state	Future state
Security and compliance	Movement and copying of data leave it susceptible to data breaches.	TREs and federation provide a secure way to access genomic data - data is not moved or copied.
Data size and interoperability	Traditional on-premise approaches are limited by scalability, cost and efficiency. Without standardised formats and pipelines, data is not interoperable.	Fully standardised data, securely accessible by cloud-based platforms, can be combined with global cohorts and disparate datasets.
Collaboration and insights	Data cannot leave jurisdictional borders and researchers cannot easily establish data sharing agreements, collaboration and insights are hindered.	No/low-code tools will support a wider range of end users enabling full democratisation of access to genomic data. Federated approaches will improve the ability to collaborate across global datasets.

Conclusion

In summary, for organisations to future-proof their approach to genomic data management, they should consider the increasing size and complexity of genomic data and growing regulations surrounding data movement. Moving towards a federated, low code, end-to-end solution for secure genomic data access will help ensure their approach is effectively future-proofed

By taking full advantage of advances in data standardisation, cloud computing, federated analysis and end-to-end data management platforms, research institutions, healthcare systems and genomic medicine programs globally can harness the benefits of collaboration and joint analyses.

This can be achieved by connecting more and diverse datasets to democratise access to data and insights, while ensuring that data stewards retain autonomy over data. In turn this will facilitate benefits sharing and therefore promote equitable access to data and clinical insights as well as international scientific collaboration.

References

1. Stark, Z. *et al.* **Integrating Genomics into Healthcare: A Global Responsibility**. *Am. J. Hum. Genet.* **104**, 13–20 (2019).

2. Macken, W. L. *et al.* **Specialist multidisciplinary input maximises rare disease diagnoses from whole genome sequencing**. *Nat. Commun.* **13**, 6324 (2022).

3. Foss, K. S. *et al.* **The Rise of Population Genomic Screening: Characteristics of Current Programs and the Need for Evidence Regarding Optimal Implementation**. *J. Pers. Med.* **12**, (2022).

4. Lander, E. S. *et al.* **Initial sequencing and analysis of the human genome.** *Nature* **409**, 860–921 (2001).

5. Berger, B. & Yu, Y. W. **Navigating bottlenecks and trade-offs in genomic data analysis.** *Nat. Rev. Genet.* (2022) doi:10.1038/s41576-022-00551-z.

6. Hasin, Y., Seldin, M. & Lusis, A. **Multi-omics approaches to disease.** *Genome Biol.* **18**, 83 (2017).

7. Schatz, M. C. *et al.* **Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space.** *Cell Genomics* **2**, 100085 (2022).

8. Alvarellos, M. *et al.* **Democratizing clinical-genomic data: How federated platforms can promote benefits sharing in genomics.** *Front. Genet.* **13**, (2023).

9. Farnaes, L. *et al.* **Rapid whole-genome sequencing decreases infant morbidity and cost of hospitalization.** *Npj Genomic Med.* **3**, 10 (2018).

10. Birney, E., Vamathevan, J. & Goodhand, P. **Genomics in healthcare: GA4GH looks to 2022.** (2017) doi:10.1101/203554.

11. Visscher, P. M. *et al.* **10 Years of GWAS Discovery: Biology, Function, and Translation**. *Am. J. Hum. Genet.* **101**, 5–22 (2017).

12. Grossman, R. L. **Data Lakes, Clouds, and Commons: A Review of Platforms for Analyzing and Sharing Genomic Data.** *Trends Genet.* **35**, 223–234 (2019).

13. Chaterji, S. *et al.* **Federation in genomics pipelines: techniques and challenges.** *Brief. Bioinform.* **20**, 235–244 (2019).

14. Rehm, H. L. *et al.* **GA4GH: International policies and standards for data sharing across genomic research and healthcare.** *Cell Genomics* **1**, 100029 (2021).

15. The Galaxy Community. **The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update.** *Nucleic Acids Res.* **50**, W345–W351 (2022).

16. Kullo, I. J., Jarvik, G. P., Manolio, T. A., Williams, M. S. & Roden, D. M. **Leveraging the electronic health record to implement genomic medicine.** *Genet. Med.* **15**, 270–271 (2013).

17. Lau-Min, K. S. *et al.* **Real-world integration of genomic data into the electronic health record: the PennChart Genomics Initiative.** *Genet. Med.* **23**, 603–605 (2021).

lifebit

lifebit.ai

Get in touch:

Request a Platform Demo

Email us at hello@lifebit.ai

At Lifebit we provide data standardisation and offer secure, federated, end-to-end TRE solutions for clients including Genomics England, Cambridge Biomedical Research Centre, Danish National Genome Centre and Boehringer Ingelheim to help researchers turn data into discoveries.

Genomics England

DANISH NATIONAL GENOME CENTER

UNIVERSITY OF CAMBRIDGE

Boehringer Ingelheim