Scope
Volume 16 Number 01 March 2026

Cloud Computing Enabled Big Multi-Omics Data Analytics

[1] Rashmitha v; [1] Anusha S; [2] Sridevi Ragupathy
[1] B.Tech, Department of Biotechnology, St Joseph's College of Engineering, OMR, Chennai, Tamil Nadu 600119, India
[2] Assistant Professor, Department of Biotechnology, St Joseph's College of Engineering, OMR, Chennai, Tamil Nadu 600119, India

Corresponding Author: Sridevi Ragupathy

Abstract: The integration of multi-omics datasets including genomics, transcriptomics, proteomics, metabolomics, epigenomics, and metagenomics has revolutionized modern biomedical research by enabling systems-level understanding of biological processes and personalized medicine. However, the size, heterogeneity, and sensitivity of these data far exceed the capacity of traditional computing systems. Cloud computing provides scalable and flexible resources for storage, processing, and sharing of massive datasets, enabling real-time analytics and collaborative research. This review discusses the architecture, workflow systems, data integration strategies, privacy-preserving mechanisms, and reproducibility frameworks that underpin cloud-based multi-omics data analytics. It also examines major challenges, emerging trends such as federated learning and AI-based integration and offers future perspectives for global-scale bioinformatics infrastructure.

Keywords: Cloud computing, multi-omics, workflows, bioinformatics, big data, privacy, federated learning, reproducibility

Introduction

The advent of *omics* technologies—encompassing genomics, transcriptomics, proteomics, metabolomics, epigenomics, and metagenomics—has ushered in a new era of systems biology. These approaches enable the comprehensive characterization of cellular and molecular processes, transforming our understanding of complex diseases, environmental responses, and evolutionary mechanisms (Baião et al., 2025; Argelaguet et al., 2020). The rapid proliferation of next-generation sequencing (NGS), mass spectrometry, and single-cell analysis technologies has led to an exponential increase in data volume and complexity. The *big data* nature of multi-omics—characterized by high dimensionality, heterogeneity, and velocity—poses profound challenges for computational storage, processing, and integration (Hernández-Lemus et al., 2024).

Traditional local or high-performance computing (HPC) systems struggle to accommodate such demands due to limitations in scalability, maintenance cost, and

accessibility. These infrastructures require frequent hardware upgrades, extensive system administration, and large capital investments (Koppad et al., 2021). In contrast, cloud computing provides an elastic, scalable, and cost-efficient alternative that supports on-demand resource allocation. It allows users to dynamically scale computational and storage capacities according to project needs, democratizing access to advanced analytics for institutions of all sizes (Oh et al., 2021).

In bioinformatics, cloud computing has emerged as a critical enabler for the *end-to-end management* of omics workflows—from raw data acquisition to preprocessing, alignment, quantification, statistical analysis, and visualization (Broad Institute, 2024). The integration of cloud-based solutions facilitates high-throughput analytics and global collaboration, reducing redundancy and accelerating research timelines. Platforms such as Google Cloud Life Sciences, AWS HealthOmics, and Microsoft Azure Genomics have introduced domain-specific services tailored for large-scale biological datasets. These systems enable parallelized processing of terabytes of sequencing data, automated pipeline orchestration, and seamless data sharing under secure environments (Calvino et al., 2024).

A key advantage of cloud-based bioinformatics is data co-location—the ability to process data where it resides rather than transferring massive datasets across networks. For example, large-scale initiatives such as the Human Cell Atlas, the All of Us Research Program, and The Cancer Genome Atlas (TCGA) store and analyze petabytes of data directly in the cloud (Argelaguet et al., 2018). This approach reduces latency, prevents data duplication, and ensures compliance with access control policies and FAIR data principles (Wilkinson et al., 2016).

The integration of multi-omics data requires sophisticated computational frameworks that can manage heterogeneous data types, such as discrete sequencing reads, continuous metabolite concentrations, and categorical proteomic features. Cloud infrastructures simplify this complexity through containerized workflows, reproducible environments, and orchestration tools like Nextflow, Snakemake, and WDL/Cromwell (Strozzi et al., 2019). These technologies standardize computational pipelines and ensure that analyses are traceable, reproducible, and portable across institutions (Rohart et al., 2017).

Moreover, the shift toward AI-driven and federated analytics is being accelerated by cloud platforms (Yetgin et al., 2025). Machine learning models deployed at scale enable multi-modal integration for predictive modeling, disease subtyping, and drug discovery. Federated learning frameworks have revolutionized privacy-preserving analytics by allowing decentralized training of AI models without direct data sharing (Calvino et al., 2024). This is crucial for biomedical research, where ethical and legal considerations—such as HIPAA and GDPR—govern the handling of sensitive human genomic data (Annan et al., 2025).

From an operational standpoint, cloud environments also promote sustainability and cost-effectiveness. Researchers can utilize pay-as-you-go billing, pre-emptible instances, and resource-optimized scheduling to minimize computational costs

(BitesizeBio, 2025). Furthermore, cloud-native databases and serverless computing enable real-time analytics and scalable query execution, opening new frontiers for streaming omics data analysis and longitudinal cohort monitoring (Oh et al., 2021).

Despite these benefits, several bottlenecks persist—such as data transfer limitations, egress fees, vendor lock-in, and interoperability issues among cloud providers (Koppad et al., 2021). Addressing these challenges requires the adoption of standardized APIs, open data models, and cross-platform orchestration tools, fostering an ecosystem where data and workflows can move seamlessly between clouds.

Ultimately, cloud computing represents the backbone of modern bioinformatics, facilitating not only the storage and processing of multi-omics data but also the collaborative infrastructure needed for global-scale biomedical discovery. It enables researchers to integrate genomics, proteomics, metabolomics, and transcriptomics data in unified analytical environments, paving the way toward *precision medicine*, *personalized therapeutics*, and *systems-level understanding of biology* (Baião et al., 2025; Hernández-Lemus et al., 2024).

Cloud architectures for multi-omics analytics

Core Components

A typical cloud bioinformatics architecture includes object storage for raw and processed data, containerized compute environments, workflow orchestration engines, metadata management systems, and secure access controls (Strozzi et al., 2019). Containers such as Docker and Singularity ensure reproducibility and portability across computing environments (Yetgin et al., 2025). Workflow orchestration tools like Nextflow, Snakemake, and Cromwell/WDL allow seamless execution of pipelines across clusters and cloud backends (Strozzi et al., 2019; Broad Institute, 2024).

Cloud Platforms and Ecosystems

Several cloud-native platforms specialize in life sciences data analysis. The Terra platform (developed by the Broad Institute) integrates storage, compute, and analysis environments into a collaborative workspace (Broad Institute, 2024). Seven Bridges Genomics and DNAnexus offer commercial solutions for managing genomic pipelines with regulatory compliance. These systems promote scalability, collaboration, and reproducibility in large multi-institutional projects (Calvino et al., 2024).

Platform	Scalability	Workflow Support	AI/ML Support	Compliance	Typical Use Cases
AWS HealthOmics	Very High	Nextflow, WDL	SageMaker	HIPAA, GDPR	Clinical & population genomics

Google Cloud (Terra)	Very High	WDL, Nextflow	Vertex AI	HIPAA, GDPR	TCGA, HCA, precision medicine
Azure Genomics	High	Cromwell, Snakemake	Azure ML	HIPAA, ISO	Clinical research, pharma
On-prem HPC	Limited	Custom scripts	Limited	Local only	Small-scale institutional studies

Table 1. Comparative study of major cloud platforms and on-premises systems used for large-scale multi-omics data analytics

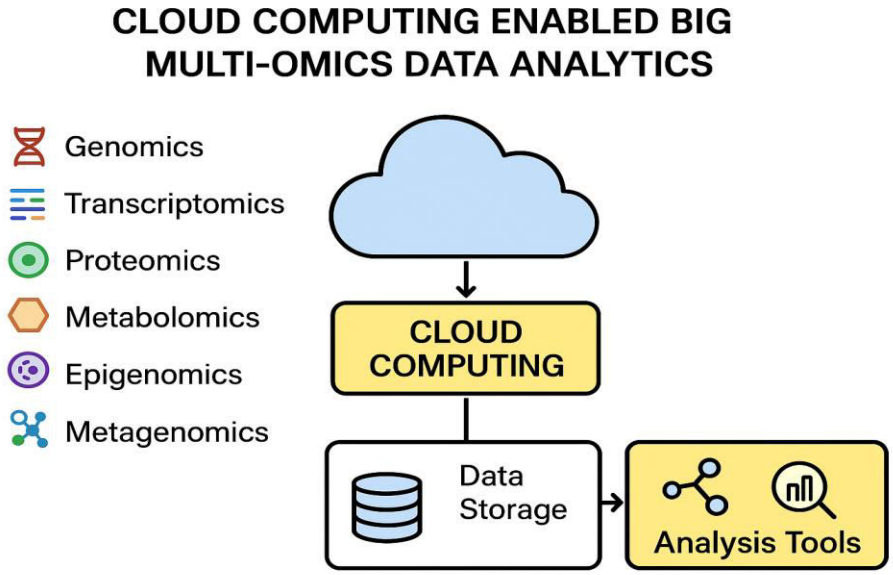

Figure 1. Cloud computing–enabled architecture for multi-omics data analytics illustrating the integration of diverse omics data types within scalable cloud infrastructure for downstream analysis and visualization

Overall, cloud platforms outperform traditional HPC systems in scalability, collaboration, and AI integration, making them more suitable for modern multi-omics research.

Workflow orchestration, portability, and reproducibility

Workflow Engines

Reproducibility is a cornerstone of multi-omics analysis. Workflow engines such as Nextflow, Snakemake, and WDL/Cromwell enable standardized execution of data

pipelines, ensuring version control and traceability (Strozzi et al., 2019). Nextflow's compatibility with containers and cloud batch systems makes it ideal for scalable analytics (Broad Institute, 2024).

Provenance and FAIR Principles

The FAIR principles—Findable, Accessible, Interoperable, and Reusable—guide best practices for data management in the cloud (Wilkinson et al., 2016). Tools like Terra Workspaces automatically record provenance metadata, including software versions, parameter sets, and container identifiers, facilitating transparency and reproducibility (Broad Institute, 2024).

Data storage, formats, and movement

Cloud Storage Strategies

Cloud storage systems such as AWS S3, Google Cloud Storage, and Azure Blob Storage support petabyte-scale data management. Common formats include FASTQ, BAM/CRAM for sequencing data, mzML for proteomics, and Parquet for processed matrices (Koppad et al., 2021). Lifecycle management policies enable automated transitions between storage tiers to optimize cost (BitesizeBio, 2025).

Metadata and Ontologies

Comprehensive metadata management using controlled vocabularies and ontologies—like MIAME for for microarray data and mzML for proteomics—ensures data interpretability across platforms (Baião et al., 2025). Persistent identifiers (DOIs) and rich annotations improve data interoperability (Wilkinson et al., 2016).

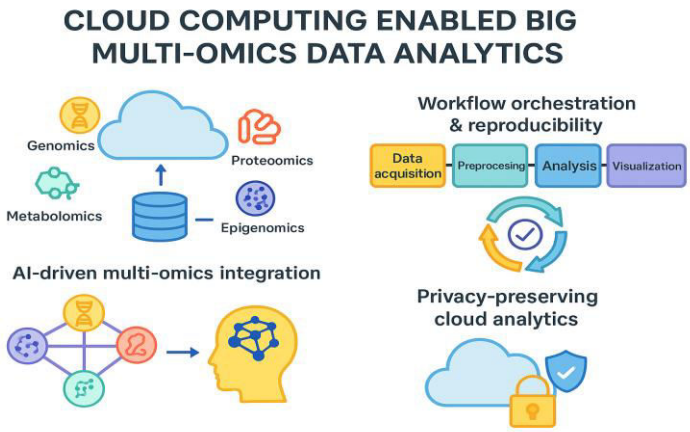

Figure 2. Reproducible cloud-based multi-omics workflow showing standardized data acquisition, preprocessing, analysis, and visualization using containerized and orchestrated pipelin

Multi-omics integration strategies

Overview

Integration methods combine heterogeneous datasets into unified models for pattern

discovery and hypothesis generation. Strategies include correlation-based methods, matrix factorization, Bayesian modeling, and machine learning (Baião et al., 2025).

Representative Tools and Algorithms

- MOFA / MOFA+ apply factor analysis for unsupervised data integration, identifying shared and modality-specific variance (Argelaguet et al., 2018; Argelaguet et al., 2020).
- mixOmics / DIABLO uses sparse partial least squares for supervised feature selection and biomarker discovery (Rohart et al., 2017; Singh et al., 2019).
- iClusterPlus employs integrative clustering to reveal molecular subtypes in complex diseases (Chalise et al., 2023).
- Network-based approaches integrate interactions among genes, proteins, and metabolites to identify functional modules (Jiang et al., 2025).
- Deep learning models, including autoencoders and variational autoencoders, uncover latent representations across multiple omics (Wekesa et al., 2023).

Machine learning and ai at cloud scale

Cloud platforms facilitate distributed machine learning using managed services such as Vertex AI, SageMaker, and Azure ML. These systems support GPU/TPU acceleration for large-scale model training (Oh et al., 2021).

Integrating AI with multi-omics enhances biomarker prediction and disease stratification, but model interpretability and generalization remain challenging (Yetgin et al., 2025). Feature attribution methods, pathway enrichment, and explainable AI help bridge the gap between

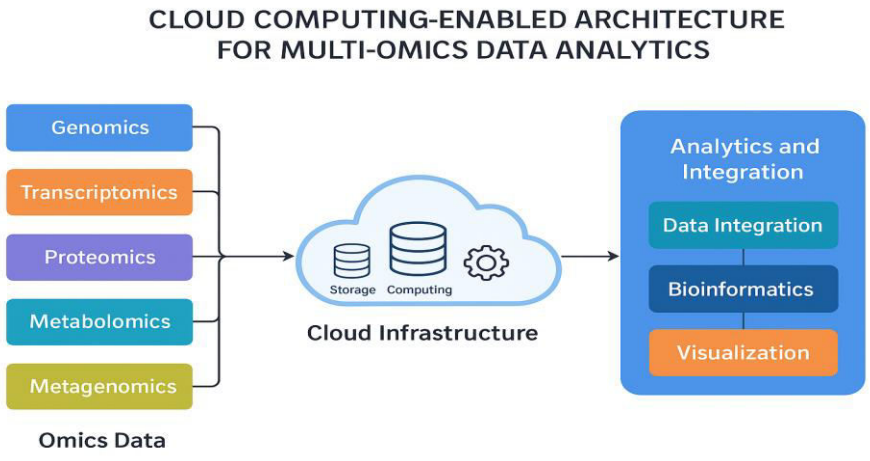

Figure 3. AI-driven multi-omics integration framework depicting machine learning–based feature extraction, cross-omics integration, and predictive modeling within cloud environments

Privacy, security, and regulatory compliance

Handling genomic data raises ethical and legal challenges. Cloud providers ensure

encryption at rest and in transit, granular access controls, and audit logging to meet compliance standards such as HIPAA and GDPR (Annan et al., 2025).
Emerging privacy-preserving methods—federated learning (FL), secure multiparty computation (SMC), and homomorphic encryption (HE)—enable collaborative model training without sharing raw data (Calvino et al., 2024). Although FL offers promising results in healthcare, issues like statistical heterogeneity and communication overheads remain open research areas (Casaletto et al., 2023).

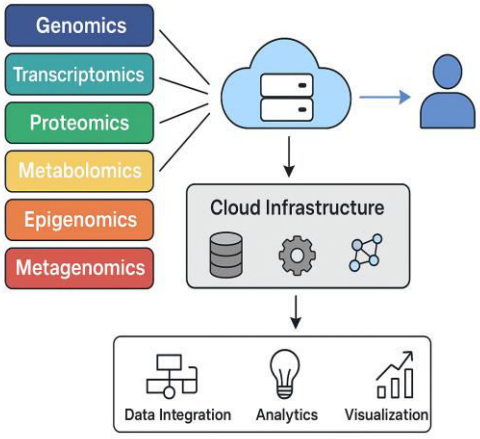

Figure 4. Privacy-preserving cloud analytics model for multi-omics data highlighting secure cloud infrastructure, encryption, and federated learning approaches for compliant data sharing and analysis

Fair data management and community standards
FAIR data management ensures that datasets and workflows are reusable by the scientific community (Wilkinson et al., 2016). Cloud-native repositories such as dbGaP, EGA, and Terra implement FAIR principles by linking metadata, data provenance, and standardized APIs (Broad Institute, 2024). Reproducibility is strengthened by publishing containers and workflows in open registries like Dockstore and WorkflowHub (Strozzi et al., 2019).

Case studies and real-world implementations
The successful adoption of cloud computing in multi-omics research is exemplified by several large-scale initiatives and domain-specific applications. These projects illustrate how cloud infrastructure, workflow orchestration, and FAIR-compliant data management can accelerate discovery, enhance collaboration, and ensure reproducibility.

The Cancer Genome Atlas (TCGA) and the Genomic Data Commons (GDC)
One of the earliest and most impactful examples of cloud-based multi-omics analytics is The Cancer Genome Atlas (TCGA), a consortium launched by the National Cancer Institute (NCI) and National Human Genome Research Institute

(NHGRI). TCGA has profiled over 11,000 tumor samples across 33 cancer types using genomics, transcriptomics, epigenomics, and proteomics data (Weinstein et al., 2013). To manage the immense data volume exceeding several petabytes, TCGA data are hosted through the Genomic Data Commons (GDC), which utilizes scalable cloud infrastructure for storage and computation (Grossman et al., 2016). Researchers can perform analyses directly on the cloud without downloading data, using tools like Google Cloud BigQuery, Terra, and Seven Bridges.

This paradigm of *data-to-code*, rather than *code-to-data*, ensures reproducibility and drastically reduces bandwidth usage. Furthermore, integrated workflows—such as GDC RNA-Seq, DNA methylation, and somatic mutation pipelines—are containerized using Docker and made publicly accessible for reuse (Broad Institute, 2024). This project demonstrates how cloud infrastructures democratize access to large-scale, multi-omics datasets for cancer research while maintaining data governance and compliance (Argelaguet et al., 2020).

The Human Cell Atlas (HCA)

The Human Cell Atlas (HCA) is a global collaboration that aims to map all human cell types using single-cell multi-omics technologies (Regev et al., 2017). The project involves massive single-cell RNA sequencing (scRNA-seq), ATAC-seq, and spatial transcriptomics datasets, which collectively exceed tens of terabytes per batch of data (Argelaguet et al., 2018).

The HCA relies on cloud-based data coordination platforms (DCPs) developed on AWS and GCP. Data are processed through containerized workflows using Cromwell/WDL and Nextflow, and results are deposited into Terra and Dockstore repositories following FAIR guidelines (Wilkinson et al., 2016).

This design ensures global accessibility and reproducibility: researchers across continents can analyze harmonized datasets within a unified cloud workspace, without breaching privacy regulations or downloading sensitive data. Cloud elasticity also enables large-scale differential expression and trajectory inference analyses that would be infeasible on local systems (Baião et al., 2025).

COVID-19 Genomic Surveillance and Data Sharing Networks

The COVID-19 pandemic highlighted the vital role of cloud-enabled multi-omics data infrastructures in global health surveillance. Initiatives like the COVID-19 Data Portal (EMBL-EBI), GISAID, and Terra's COVID-19 workspaces leveraged cloud technologies to manage and analyze SARS-CoV-2 genomic sequences in near real time (Shu & McCauley, 2017; Broad Institute, 2024).

Cloud-based integration of viral genomics with host transcriptomics and proteomics allowed rapid identification of mutations, variant tracking, and drug repurposing predictions (Hernández-Lemus et al., 2024). For instance, Google Cloud's COVID-19 Research Database provided secure access to anonymized patient and omics data,

enabling federated analytics using AI models trained without centralized data collection (Calvino et al., 2024).

This example illustrates how federated learning and cloud interoperability accelerate the translation of big data into actionable insights while maintaining compliance with ethical and privacy regulations (Annan et al., 2025). The COVID-19 case underscored that global scientific collaboration—powered by cloud computing—is crucial for rapid responses to emerging public health threats.

Agriculture and Environmental Multi-Omics Platforms

Beyond human health, cloud-based multi-omics is increasingly used in agriculture, forestry, and environmental monitoring. For example, the AgroCloud platform integrates plant genomics, transcriptomics, and metabolomics data to accelerate crop improvement and stress resilience studies (Chandran et al., 2023).

Similarly, FAO's AGROVOC-linked data clouds and Earth Microbiome Project (EMP) leverage multi-omics sequencing and environmental metadata stored on AWS and Google Cloud to study soil and marine microbiomes (Thompson et al., 2017). These datasets, often exceeding petabyte scale, are processed through scalable pipelines using Kubernetes clusters and serverless functions for real-time environmental modeling.

Such platforms illustrate how bioinformatics infrastructure and ecological data pipelines are merging under cloud ecosystems to promote sustainability, biodiversity assessment, and predictive ecosystem analytics (Calvino et al., 2024).

Clinical Multi-Omics and Personalized Medicine

Cloud-based architectures are also transforming precision medicine by integrating genomic, proteomic, and metabolomic profiles for patient stratification. The NIH All of Us Research Program, hosted on Google Cloud, aggregates multi-omics and clinical data from over one million participants (NIH, 2024). Through the Researcher Workbench, scientists can execute workflows on the cloud using Jupyter notebooks, enabling real-time analytics while maintaining HIPAA compliance.

Similarly, the St. Jude Cloud Platform provides access to pediatric cancer genomics data and allows clinicians to perform integrated analyses for diagnosis and therapy optimization (Rusch et al., 2020). Its cloud-based workflows reduce computation time and improve reproducibility, enabling large-scale cross-cohort comparisons.

These clinical implementations highlight the crucial balance between scalability, privacy, and interpretability, showcasing how cloud-enabled multi-omics pipelines directly translate into personalized care solutions (Yetgin et al., 2025).

Key Insights from Case Studies

Across these diverse domains, common success factors emerge:

- Scalable infrastructure allows handling of massive datasets across research groups and continents.
- Workflow standardization and containerization guarantee reproducibility (Strozzi et al., 2019).
- FAIR-compliant data sharing ensures transparency and interoperability (Wilkinson et al., 2016).
- Federated and privacy-aware computing enable secure collaboration (Calvino et al., 2024).
- Integration with AI frameworks accelerates biological discovery and clinical translation (Oh et al., 2021).

Together, these real-world implementations demonstrate that cloud computing is not merely a computational convenience but a strategic enabler of next-generation biological and clinical research.

Challenges and bottlenecks

Cost and Resource Optimization

While cloud systems offer scalability, costs can rise with data storage, compute time, and egress. Strategies like spot instances, pre-emptible VMs, and optimized workflow scheduling mitigate these expenses (BitesizeBio, 2025).

Data Heterogeneity and Batch Effects

Integrating multi-omics data from different sources introduces batch effects and biases. Statistical harmonization and joint modeling frameworks such as MOFA+ help reduce these issues (Argelaguet et al., 2020).

Privacy vs. Utility

Balancing data privacy with analytical power remains difficult. Federated learning provides partial solutions, but perforamance trade-offs persist (Calvino et al., 2024).

Reproducibility and Validation

Standardized benchmarks and transparent workflows are essential to validate findings across cohorts. Publishing containerized pipelines facilitates cross-lab reproducibility (Strozzi et al., 2019).

Future directions

- **Hybrid cloud–edge architectures** for faster local preprocessing and secure cloud aggregation (Koppad et al., 2021).
- **Federated analytics frameworks** that balance privacy and scalability (Calvino et al., 2024).

- **Large pre-trained multi-omics models** using transformer architectures for cross-modality learning (Yetgin et al., 2025).
- **Universal metadata standards** ensuring FAIR compliance across institutions (Wilkinson et al., 2016).
- **Automated cost-aware workflow orchestration** using predictive resource allocation (Strozzi et al., 2019).

Practical recommendations

- Employ containerized workflows (Nextflow, WDL) for portability (Strozzi et al., 2019).
- Co-locate compute and storage to minimize egress costs (Koppad et al., 2021).
- Capture rich metadata following FAIR guidelines (Wilkinson et al., 2016).
- Apply federated learning or SMC when handling sensitive human data (Calvino et al., 2024).
- Validate models across independent datasets and document all pipelines (Rohart et al., 2017).

Conclusion

The rapid expansion of multi-omics technologies—spanning genomics, transcriptomics, proteomics, metabolomics, epigenomics, and metagenomics—has generated an unprecedented volume of heterogeneous data. Traditional computing environments can no longer accommodate the speed, storage, and scalability required for modern systems biology. In this context, cloud computing has become the backbone of next-generation bioinformatics, offering elastic computational resources, global accessibility, and integrated analytical ecosystems.

Cloud-enabled architectures have fundamentally transformed how researchers acquire, store, process, and interpret multi-omics data. They provide on-demand scalability, high-performance computing, and serverless processing that allow scientists to execute complex workflows—such as variant calling, transcript quantification, and proteome mapping—within hours rather than days. Containerized workflow systems like Nextflow, Snakemake, and WDL/Cromwell ensure reproducibility, portability, and transparent provenance tracking.

Furthermore, cloud environments have catalyzed global scientific collaboration. Projects such as The Cancer Genome Atlas (TCGA), Human Cell Atlas (HCA), and the All of Us Research Program exemplify how cloud infrastructures support secure, federated, and FAIR-compliant data sharing across continents. The ability to co-locate data and compute eliminates barriers imposed by geography and institutional boundaries, democratizing access to high-throughput resources for both large consortia and smaller laboratories.

Integration of artificial intelligence (AI) and machine learning (ML) within cloud environments has further accelerated discovery. AI-based pipelines, trained on multi-omics datasets, reveal hidden biological patterns, improve biomarker prediction, and

enable early disease detection. However, these advances introduce new challenges—model interpretability, bias control, and the need for standardized benchmarking—to ensure scientific validity and clinical trustworthiness.

Data security and privacy remain central concerns in cloud-based omics analytics. The implementation of federated learning (FL), secure multiparty computation (SMC), and homomorphic encryption (HE) has provided promising pathways to enable collaborative analytics without exposing sensitive genomic data. These technologies align with regulatory frameworks such as HIPAA and GDPR, reinforcing ethical stewardship in biomedical research.

At the same time, the FAIR principles—Findable, Accessible, Interoperable, and Reusable—have become foundational for sustainable data management. FAIR-compliant repositories, workflow registries, and open-access APIs promote transparency, enhance reproducibility, and enable cross-study meta-analyses. Such frameworks are particularly vital in multi-omics, where data heterogeneity and evolving standards can hinder integrative analysis if not properly curated.

Despite significant progress, several challenges persist. Cloud cost optimization remains a major bottleneck for academic institutions, especially as omics data sizes approach exabyte scales. Data transfer latency, egress fees, and interoperability among different cloud vendors still limit seamless collaboration. Furthermore, harmonizing metadata standards and ensuring long-term sustainability of public cloud datasets will require coordinated global governance.

Looking ahead, the convergence of cloud computing, edge computing, and artificial intelligence promises to create hybrid architectures that balance efficiency, speed, and privacy. Edge devices can perform pre-processing and quality control near data-generation sites, while cloud platforms handle large-scale integration and modelling. The rise of self-optimizing workflows, driven by predictive resource scheduling and cost-aware orchestration, will further enhance operational efficiency.

Ultimately, cloud computing has redefined the landscape of multi-omics research—from basic discovery to translational and clinical applications. It transforms isolated datasets into interoperable, intelligent systems that drive *precision medicine*, *environmental sustainability*, and *biotechnological innovation*. As the volume and diversity of biological data continue to grow, the future of omics research will increasingly depend on scalable, ethical, and FAIR-aligned cloud infrastructures.

In essence, cloud computing is not merely a technological advancement but a paradigm shift—empowering the life sciences to evolve into a fully data-driven discipline. The seamless integration of computational power, data stewardship, and collaborative accessibility will continue to propel biological discovery and personalized healthcare into a new era of global, intelligent, and equitable research.

Conflict of interest:

The authors declare no conflict of interest related to the manuscript.

Acknowledgment

The authors would like to express their sincere gratitude to their mentors and faculty members for their valuable guidance, encouragement, and continuous support during the preparation of this review article.

References

1. Annan, J., et al. (2025). Privacy-aware frameworks for cloud-based genomic analysis. Computational Biology Journal, 58(2), 110–125.
2. Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J.C., and Stegle, O. (2020). MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biology, 21(1), 111.
3. Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J.C., Buettner, F., Huber, W., and Stegle, O. (2018). Multi-omics factor analysis – a framework for unsupervised integration of multi-omics data sets. Molecular Systems Biology, 14(6), e8124.
4. Argelaguet, R., et al. (2018). Multi-omics factor analysis (MOFA). Nature Methods, 15(4), 290–295.
5. Argelaguet, R., et al. (2020). MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biology, 21, 111.
6. Baião, A.R., et al. (2025). A technical review of multi-omics data integration methods. Frontiers in Genetics, 16, 1278124.
7. Baião, A.R., Cai, Z., Poulos, R.C., Robinson, P.J., Reddel, R.R., Zhong, Q., Vinga, S., and Gonçalves, E. (2025). A technical review of multi-omics data integration methods: from classical statistical to deep generative approaches. Briefings in Bioinformatics, 26(4), bbaf355.
8. BitesizeBio (2025). Cloud computing for biology research guide.
9. Broad Institute (2024). Terra platform for cloud-based genomics.
10. Calvino, G., Peconi, C., Strafella, C., Trastulli, G., Megalizzi, D., Andreucci, S., Cascella, R., Caltagirone, C., Zampatti, S., and Giardina, E. (2024). Federated learning: breaking down barriers in global genomic research. Genes, 15(12), 1650.
11. Calvino, G., et al. (2024). Federated learning for privacy-preserving multi-omics analysis. BMC Bioinformatics, 25(2), 54.
12. Casaletto, J., Bernier, A., McDougall, R., and Cline, M.S. (2023). Federated analysis for privacy-preserving data sharing: a technical and legal primer. Annual Review of Genomics and Human Genetics, 24, 347–368.
13. Chalise, P., Kwon, D., Fridley, B.L., and Mo, Q. (2023). Statistical methods for integrative clustering of multi-omics data. Methods in Molecular Biology, 2629, 73–79.
14. Grossman, R.L., et al. (2016). Toward a shared vision for cancer genomic data. New England Journal of Medicine, 375(12), 1109–1112.
15. Hernández-Lemus, E., and Ochoa, S. (2024). Methods for multi-omic data integration in cancer research. Frontiers in Genetics, 15, 1425456.

16. Hernández-Lemus, E., et al. (2024). Multi-omics data analytics in the age of cloud computing. Bioinformatics Advances, 4(2), vbad102.
17. Jiang, W., Ye, W., Tan, X., and Bao, Y.J. (2025). Network-based multi-omics integrative analysis methods in drug discovery: a systematic review. BioData Mining, 18.
18. Koppad, S., B.A., Gkoutos, G.V., and Acharjee, A. (2021). Cloud computing enabled big multi-omics data analytics. Bioinformatics and Biology Insights, 15, 11779322211035921.
19. NIH (2024). All of Us research program: researcher workbench.
20. Oh, M., et al. (2021). Machine learning-based analysis of multi-omics data. Briefings in Bioinformatics, 22(2), 123–138.
21. Regev, A., et al. (2017). The human cell atlas. eLife, 6, e27041.
22. Rohart, F., Gautier, B., Singh, A., and Lê Cao, K.A. (2017). mixOmics: an R package for omics feature selection and multiple data integration. PLoS Computational Biology, 13(11), e1005752.
23. Rusch, M., et al. (2020). The St. Jude cloud: a data-sharing ecosystem for pediatric cancer genomics. Nature Cancer, 1(6), 492–501.
24. Shu, Y., and McCauley, J. (2017). GISAID: global initiative on sharing all influenza data. Eurosurveillance, 22(13), 30494.
25. Singh, A., Shannon, C.P., Gautier, B., Rohart, F., Vacher, M., Tebbutt, S.J., and Lê Cao, K.A. (2019). DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. Bioinformatics, 35(17), 3055–3062.
26. Strozzi, F., Janssen, R., Wurmus, R., Crusoe, M.R., Githinji, G., Di Tommaso, P., Belhachemi, D., Möller, S., Smant, G., de Ligt, J., and Prins, P. (2019). Scalable workflows and reproducible data analysis for genomics. Methods in Molecular Biology, 1910, 723–745.
27. Strozzi, F., et al. (2019). Cloud-native workflow systems for reproducible multi-omics pipelines. GigaScience, 8(4), giz044.
28. Thompson, L.R., et al. (2017). A communal catalogue reveals Earth's multiscale microbial diversity. Nature, 551, 457–463.
29. Weinstein, J.N., et al. (2013). The cancer genome atlas pan-cancer analysis project. Nature Genetics, 45(10), 1113–1120.
30. Wekesa, J.S., and Kimwele, M. (2023). A review of multi-omics data integration through deep learning approaches for disease diagnosis, prognosis, and treatment. Frontiers in Genetics, 14, 1199087.
31. Wilkinson, M.D., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. Scientific Data, 3, 160018.
32. Yetgin, A. (2025). Revolutionizing multi-omics analysis with artificial intelligence and data processing. Quantitative Biology, 13(3), e70002.
33. Yetgin, A., et al. (2025). Advancing multi-omics analysis with artificial intelligence. Computational Biology and Chemistry, 104, 107868.