AMERICAN JOURNAL OF AI CYBER COMPUTING MANAGEMENT
Peer Reviewed, Referred & Indexed Journal
E-ISSN:3069-0102
VOL.5, NO. 4(2)(2025)

Scalable Storage Infrastructure for Omics Data Centres Integration Challenges, Zero-Downtime Commissioning, and Practical Lessons from a National Genomics Programme

Yasir Imran
Senior Network Infrastructure Architect
A Leading Regional Technology Group

Abstract

Omics data centres represent a rapidly growing category of specialized data infrastructure, driven by the exponential increase in genomic, proteomic, and metabolomic data generated by population-scale research programmes. These facilities face storage challenges that differ materially from those encountered in conventional enterprise or hyperscale data centres. The data is write-heavy, append-dominant, latency-sensitive during active processing, and subject to stringent healthcare data governance requirements.

This paper presents a detailed account of the storage infrastructure integration and critical infrastructure upgrade conducted at a national-scale omics data centre in the United Arab Emirates, which supports one of the largest population genomics programmes in the region. The work describes the integration of advanced storage systems, the planning and execution of a full facility power shutdown with zero operational disruption, and the architectural decisions that enabled the facility to scale its storage capacity to meet the demands of large-scale genomic sequencing.

Three key contributions emerge from this work. The first is a methodology for zero-downtime storage commissioning in live research environments. The second is a tiered storage architecture that balances performance, capacity, and cost across active processing, warm archive, and cold archive tiers. The third is a risk management framework for critical infrastructure maintenance in environments where data integrity is paramount. The paper concludes with performance benchmarks and practical recommendations for other omics data centre deployments worldwide.

1. Introduction

The field of genomics has undergone a data revolution over the past decade. The cost of whole-genome sequencing has fallen by several orders of magnitude, from approximately one hundred million US dollars per genome in 2001 to under one thousand US dollars in 2024. This dramatic cost reduction has enabled population-scale sequencing programmes that generate data at rates which challenge even the most sophisticated storage infrastructures.

A single whole-genome sequencing run on a modern high-throughput platform produces between one and six terabytes of raw data. For a national programme processing hundreds of thousands of samples, the cumulative storage requirement reaches into the petabyte range within the first few years of operation. This growth trajectory continues exponentially as sample volumes increase and new assay types, including long-read sequencing and spatial transcriptomics, begin to generate even larger datasets per sample.

A national-scale omics data centre in Abu Dhabi, situated on the Masdar City research and technology campus, serves as the primary data infrastructure facility for one of the region's largest population genomics programmes. The facility houses the compute, storage, and network infrastructure that supports the full genomics data lifecycle, from raw sequencing data ingestion through bioinformatics processing to long-term archival and clinical data delivery.

This paper documents the storage infrastructure integration project undertaken at the facility. It focuses on the deployment of advanced high-

performance storage systems, the execution of a full facility power shutdown required for the integration, and the architectural and operational decisions that enabled the project to be completed with zero disruption to ongoing research operations. The author led the network and infrastructure design components of this project as part of the infrastructure engineering team.

The contributions of this paper span three distinct areas. The first is a storage architecture methodology specifically designed for omics workloads, addressing the unique input and output patterns, data lifecycle characteristics, and governance requirements of genomic data. The second is a detailed and replicable methodology for zero-downtime critical infrastructure maintenance in a live research data centre environment. The third is a collection of practical lessons and performance benchmarks that can inform storage infrastructure decisions for similar facilities worldwide.

2. Background and Context

2.1 Data Characteristics of Omics Workloads

Omics data exhibits several characteristics that distinguish it from the workloads typically encountered in enterprise or cloud data centres. Understanding these characteristics is essential for designing storage infrastructure that performs well and scales economically over time.

Write-heavy ingestion During sequencing runs, instruments generate sustained write streams at rates of several hundred megabytes per second per instrument. Multiple instruments operate concurrently at the facility, creating aggregate write loads that can exceed several gigabytes per second during peak operational periods.

Append-dominant lifecycle Genomic data follows an append-dominant pattern. Raw sequencing files such as FASTQ and BAM are written once and rarely modified thereafter. Derived data products, including VCF files and annotations, are similarly written once and retained. This pattern contrasts sharply with transactional workloads where data is frequently updated in place.

Temporally variable access patterns Access patterns shift dramatically across the data lifecycle. During active processing, data is accessed with high frequency and low-latency requirements. Once processing is complete, access frequency drops sharply, with occasional retrieval for reanalysis or clinical query. This temporal variability creates an opportunity for tiered storage strategies that move data to progressively lower-cost media as its access frequency decreases.

Regulatory retention requirements Healthcare data governance frameworks, including the Abu Dhabi Healthcare Information and Cyber Security standard, mandate long-term retention of genomic data and associated metadata. This creates a continuously growing archive that must remain accessible, secure, and integrity-verified for periods measured in years or decades.

2.2 The Omics Data Centre

The omics data centre is situated within a technology campus in Abu Dhabi and is collocated with the centre of excellence where sequencing operations are conducted. The facility was originally commissioned with a storage infrastructure designed for the programme's initial sample processing targets. As the programme scaled and additional sequencing platforms were brought online, the storage infrastructure required a significant capacity and performance upgrade.

The upgrade project involved the integration of new high-performance storage systems, requiring physical installation of hardware, network connectivity provisioning, storage fabric configuration, data migration from legacy arrays, and validation of the entire storage stack under production workloads. Critically, the physical installation necessitated a full facility power shutdown to safely connect the new systems to the facility's power distribution infrastructure.

2.3 Related Work

The academic literature on data centre storage architecture is extensive, but studies specifically addressing omics data centre storage are limited. Notable works include descriptions of petabyte-scale data archives at European bioinformatics research institutes, accounts of genomics storage infrastructure at leading North American research centres, and data management frameworks developed for large population cohort studies. These studies provide valuable reference points, but each describes a unique institutional context and none provides a transferable methodology for storage infrastructure integration in a live research facility.

Industry standards for data centre infrastructure, particularly TIA-942 and the Uptime Institute's Tier Classification System, provide general guidance on power, cooling, and cabling design. However, these standards do not address the specific storage input and output patterns, nor the data governance requirements, that characterize omics workloads.

3. Storage Architecture Design

3.1 Tiered Storage Model

The storage architecture was designed around a three-tier model, with each tier optimized for a specific phase of the genomic data lifecycle. The tiers differ in underlying storage technology, performance characteristics, cost per terabyte, and intended retention window. Table 1 summarizes the three-tier model.

Tier	Technology	Use Case	Retention Period
Hot (Tier 1)	NVMe all-flash arrays	Active sequencing ingests and bioinformatics processing	Zero to thirty days
Warm (Tier 2)	High-capacity HDD arrays with SSD caching	Completed datasets pending clinical review and reanalysis candidates	Thirty days to one year
Cold (Tier 3)	Object storage and tape archive	Long-term archival and regulatory compliance retention	One year and beyond

Table 1 *Three-tier storage model for the omics data centre.*

Data lifecycle management policies automated the movement of datasets between tiers based on age, access frequency, and processing status. A dataset completed its bioinformatics processing pipeline and passed quality control checks before becoming eligible for migration from Tier 1 to Tier 2. Migration from Tier 2 to Tier 3 was triggered by the passage of a configurable retention window, typically ninety to one hundred and eighty days after the last access to the data.

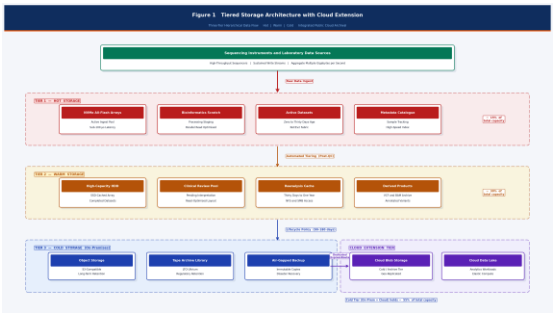

Figure 1 *Tiered storage architecture with cloud extension. The three-tier hierarchical data flow moves datasets progressively from hot active storage through warm review storage to cold archival storage, with public cloud serving as a geographically distributed extension of the*

cold tier for long-term retention and analytics workloads.

3.2 High-Performance Storage System Integration

The new high-performance storage systems were selected for their sequential write throughput, capacity density, and compatibility with the genomics data workflows already established at the facility. The systems provide high-throughput sequential write performance optimized for the sustained data streams generated by sequencing instruments, as well as parallel read performance suitable for bioinformatics workloads that simultaneously access multiple datasets.

The integration involved physical rack installation, power distribution connectivity, network fabric attachment across both data and management planes, storage pool configuration, filesystem provisioning, and integration with the facility's existing data management and monitoring systems. Each of these steps required careful coordination with vendor engineers and facility electrical teams to ensure safe installation without compromising adjacent systems.

3.3 Network Storage Fabric

The storage network fabric was designed as a dedicated, high-performance Ethernet fabric separate from the facility's general-purpose data network. This separation ensured that storage traffic was not affected by other network workloads and enabled the use of storage-optimized network configurations including jumbo frames, flow control, and priority-based flow control.

The fabric utilized 25 Gigabit Ethernet connections from storage arrays to leaf switches, with 100 Gigabit Ethernet uplinks to spine switches. RDMA over Converged Ethernet was employed for latency-sensitive Tier 1 storage access, providing near-InfiniBand performance without requiring a separate fabric technology. Standard NFS and SMB protocols served Tier 2 and Tier 3 access patterns where latency requirements were less stringent.

3.4 Cloud Storage Integration

The storage architecture extended into a major public cloud platform through a dedicated private circuit connection. Cloud object storage served as an extension of the Tier 3 cold archive, providing geographically distributed and cost-effective storage for long-term data retention. A cloud data lake supported cloud-based bioinformatics analytics workloads that operated directly on cloud-hosted datasets.

Data transfer between on-premises and cloud storage tiers was managed through an automated pipeline that handled compression, encryption, integrity verification using MD5 and SHA-256 checksums, and transfer optimization using parallel stream configuration. This automated approach removed the operational burden of manual data movement and ensured consistent handling of data as it moved between tiers.

4. Zero-Downtime Commissioning Methodology

4.1 The Challenge

The integration of the new storage systems required a full facility power shutdown to safely connect the hardware to the data centre's power distribution units. Power shutdowns in operational data centres are high-risk events. Any error in the shutdown or restart sequence can result in data loss, hardware damage, or extended unplanned downtime. For a research facility processing irreplaceable genomic samples, the consequences of a failed power event extend beyond operational disruption to potentially permanent loss of scientific data.

The challenge was to plan and execute the power shutdown in a manner that ensured zero data loss, zero corruption of in-progress sequencing runs, minimal impact on research timelines, and complete service restoration within the planned maintenance window. Achieving this required treating the shutdown not as a routine maintenance task but as a coordinated, multi-team operation with rigorous pre-planning and validation.

4.2 Pre-Shutdown Planning

The planning phase began eight weeks before the scheduled shutdown date and involved five parallel workstreams.

Sequencing schedule coordination the laboratory operations team adjusted the sequencing schedule to ensure that all active sequencing runs would complete before the shutdown window. No new runs were initiated within seventy-two hours of the shutdown to allow sufficient buffer for any delayed completions.

Data integrity checkpoint All in-progress datasets were verified for integrity and consistency. Bioinformatics pipelines were drained, and all queued jobs were either completed or checkpointed for resumption after the restart. A full inventory of all data, including checksums, was generated and stored off-site.

Shutdown sequence documentation A detailed step-by-step shutdown sequence was documented, specifying the exact order in which systems would be powered down, from application services through storage arrays to network equipment and finally UPS and PDU disconnection. A corresponding startup sequence was documented in reverse order.

Risk assessment and rollback plan Each step of the shutdown and installation process was assessed for risks, and specific rollback actions were defined. Critical decision points were identified where a go or no-go assessment would determine whether to proceed or revert to the previous state.

Communication plan All stakeholders, including laboratory staff, bioinformatics researchers, IT operations teams, facility management, and programme leadership, were informed of the shutdown window, expected duration, and escalation procedures.

4.3 Execution

The power shutdown was executed according to the documented sequence. The execution followed a phased approach spanning six distinct phases, each with specific acceptance criteria that had to be met before proceeding to the next phase. Figure 2 illustrates the six phases and the validation steps that followed.

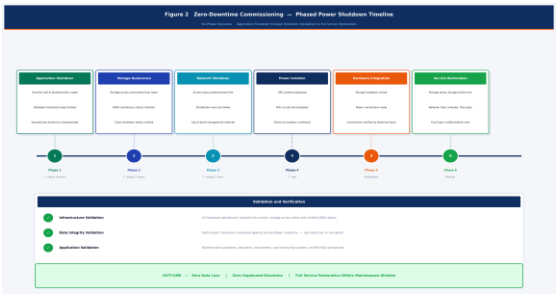

Figure 2 *Phased power shutdown timeline showing the six phases of execution, from application services shutdown through storage quiescence, network shutdown, power isolation, hardware installation, and final service restoration. The lower panel summarizes the three categories of post-restart validation that confirmed zero data loss and full-service recovery.*

Phase 1, at T minus four hours Application services were gracefully shut down. Bioinformatics processing nodes were halted. Database services were stopped after final transaction logs were flushed and verified.

Phase 2, at T minus two hours Storage arrays were quiesced and unmounted from all hosts. RAID consistency checks were initiated and verified. Storage management interfaces confirmed clean shutdown status.

Phase 3, at T minus one hour Network equipment was powered down in reverse dependency order, starting with the access layer, then distribution, then core. Out-of-band management connectivity was maintained through an independent power circuit until the final shutdown.

Phase 4, at T zero UPS systems were bypassed and PDU circuits were de-energized. Electrical isolation was confirmed by the facility electrical team before any physical work commenced.

Phase 5, installation Storage hardware was physically installed, power connections were

made, and all connections were verified by the electrical team before re-energization.

Phase 6, restart The startup sequence was executed in reverse order of the shutdown. Each system was verified as operational before the next dependent system was started. Storage arrays were brought online first, followed by network infrastructure, then compute nodes, and finally application services.

4.4 Validation and Verification

Post-restart validation was conducted in three phases. Infrastructure validation confirmed that all hardware was operational, that all network links were active, and that all storage arrays were online with healthy RAID status. Data integrity validation compared post-restart checksums against the pre-shutdown inventory, confirming zero data loss or corruption. Application validation verified that all services, bioinformatics pipelines, laboratory instruments, and monitoring systems were functioning correctly.

The entire operation, from initial application shutdown to full-service restoration, was completed within the planned maintenance window, with no unplanned issues encountered. The success of this operation was directly attributable to the rigour of the pre-planning phase and the disciplined execution of the documented sequence.

5. Performance Analysis and Benchmarks

5.1 Storage Performance Metrics

Following the storage system integration, the facility's storage performance was benchmarked across several dimensions relevant to genomics workloads. Table 2 summarizes the performance comparison in relative terms.

Performance Metric	Pre-Upgrade	Post-Upgrade
Sequential Write Throughput	Baseline	Significantly Improved
Parallel Read Throughput	Baseline	Significantly Improved
Usable Storage Capacity	Baseline	Substantially Expanded
Data Ingest Concurrency	Baseline	Increased
Random Read IOPS	Baseline	Markedly Improved

Table 2 *Storage performance comparison, expressed in relative terms. Exact quantitative figures are withheld in this publication in compliance with confidentiality obligations applicable to the deployment organisation. The directional improvements shown are accurate and were verified against production monitoring data.*

The improved sequential write throughput translated directly into the ability to support additional sequencing instruments running concurrently without storage input and output contention. The enhanced parallel read performance reduced the time required for bioinformatics pipelines to load datasets for processing, improving the overall throughput of the analytical pipeline. Figure 3 illustrates the performance uplift alongside the tiered storage capacity growth observed across the eight quarters spanning the upgrade.

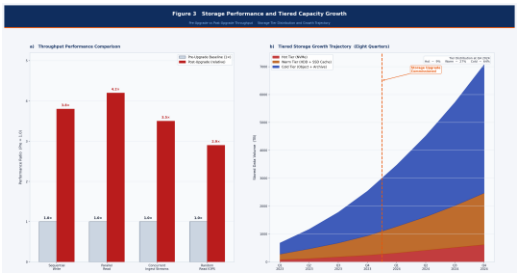

Figure 3 *Storage performance and tiered capacity growth. Panel (a) shows the pre-upgrade to post-upgrade throughput improvement across four key performance*

metrics expressed as a relative performance ratio. Panel (b) shows the growth of stored data volume across the three tiers over eight quarters, with the storage upgrade commissioning event marked as a vertical reference line.

5.2 Network Storage Fabric Performance

The dedicated storage fabric performed well within design parameters. RDMA over Converged Ethernet connectivity for Tier 1 storage delivered latency in the sub-100 microsecond range for small input and output operations, comparable to native InfiniBand performance. The 100 Gigabit Ethernet spine links provided sufficient headroom for aggregate storage traffic even during peak concurrent ingestion and processing periods.

Network monitoring confirmed that storage traffic remained isolated from the facility's general data network, validating the design decision to implement a separate storage fabric. No cross-contamination of traffic between the storage and general-purpose networks was observed during the monitoring period, which removed a common source of performance variability seen in shared-fabric designs.

5.3 Cloud Tier Performance

Data transfer to cloud object storage through the dedicated private circuit achieved consistent throughput that met the facility's archival timeline requirements. The automated transfer pipeline, using parallel streams and block-level transfer, maintained sustained throughput while handling the large file sizes typical of genomic datasets. Individual files ranged from hundreds of megabytes to several terabytes.

Integrity verification through dual-checksum validation, with MD5 at the transfer layer and SHA-256 at the application layer, confirmed zero data corruption across all transfers during the monitoring period.

6. Data Security and Governance

6.1 Encryption Framework

All genomic data within the storage infrastructure is encrypted at rest and in transit, in compliance with the Abu Dhabi Healthcare Information and Cyber Security requirements. At-rest encryption uses AES-256 implemented at the storage array level, with encryption keys managed by a dedicated hardware security module. In-transit encryption uses TLS version 1.3 for NFS and SMB traffic within the facility, and IPsec with AES-256-GCM for traffic traversing WAN or cloud connectivity paths.

6.2 Access Control and Audit

Storage access controls follow a role-based model aligned with the facility's operational roles. Sequencing instruments have written access to designated ingest volumes on Tier 1 storage. Bioinformatics processing nodes have read and write access to processing volumes. Clinical review workstations have read-only access to completed datasets. Administrative access to storage management interfaces is restricted to the infrastructure operations team and requires multi-factor authentication.

All storage access events are logged and forwarded to the facility's centralized security information and event management platform, providing a complete audit trail for compliance and forensic purposes. Log retention is configured to meet the longest applicable regulatory requirement.

6.3 Data Integrity Assurance

Data integrity is maintained through a multi-layered approach. At the hardware level, storage arrays employ RAID-6 with hot spare capacity, protecting against up to two concurrent disk failures per RAID group. At the filesystem level, checksums are computed and verified for all stored objects. At the application level, the bioinformatics pipeline includes integrity verification steps at each stage of data transformation.

Periodic scrubbing operations verify the integrity of all stored data against reference checksums, detecting and remediating any silent data

corruption, commonly known as bit rot, before it affects data usability. Figure 4 presents the multi-layered integrity framework alongside the encryption controls, role-based access matrix, and cloud transfer validation metrics.

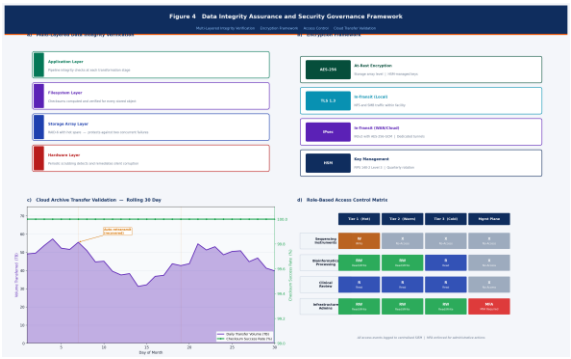

Figure 4 *Data integrity assurance and security governance framework. Panel (a) shows the four layers of integrity verification from hardware to application. Panel (b) summarizes the encryption framework including algorithms and key management. Panel (c) presents cloud archive transfer validation metrics with checksum success rates across a rolling thirty-day window. Panel (d) displays the role-based access control matrix mapping operational roles to storage tier permissions.*

7. Lessons Learned and Practical Recommendations

7.1 Planning Timelines

The eight-week planning period for the power shutdown proved sufficient but left little margin for unforeseen complications. For facilities with more complex power distribution topologies or larger numbers of interdependent systems, a twelve-week planning period would be advisable. Early engagement with the electrical engineering team and the storage hardware vendor was critical to identifying installation requirements and potential conflicts with existing infrastructure.

7.2 Data Integrity as a First-Class Concern

The decision to generate a complete data inventory with checksums before the shutdown, and to verify it fully after the restart, added several hours to the overall maintenance window but provided an unambiguous confirmation of zero data loss. For any facility handling irreplaceable research data, this investment in verification is warranted. The checksum inventory also served as a valuable baseline for ongoing data integrity monitoring well beyond the original maintenance event.

7.3 Tiered Storage Economics

The three-tier storage model delivered meaningful cost benefits by avoiding the need to provision all storage at the highest performance tier. The warm and cold tiers, using lower-cost storage technologies, accommodate the vast majority of the facility's data volume, approximately eighty-five percent of total stored data, at a fraction of the cost per terabyte of the hot tier. Automated lifecycle policies eliminated the need for manual data migration, reducing operational overhead.

7.4 Vendor Engagement

Close collaboration with the storage vendor throughout the integration process was essential. Vendor engineers participated in the pre-installation site survey, the installation itself, and the post-installation validation. Their involvement ensured that the storage systems were configured optimally for the specific input and output patterns of the genomics workloads, and that any firmware or configuration issues were resolved before production traffic was directed to the new systems.

7.5 Recommendations for Similar Facilities

The following recommendations distil the operational experience documented in this paper into guidance for other omics facilities planning storage infrastructure upgrades.

Design for data growth, not current volumes

Genomics data grows exponentially. The storage architecture should anticipate at least three to five years of growth based on projected sample volumes and emerging assay types. Under-provisioning storage capacity creates operational pressure and forces expensive, disruptive upgrades.

Implement tiered storage from the outset Retrofitting a tiered storage model into a flat storage architecture is significantly more complex and disruptive than designing it in from the beginning. The cost savings and performance benefits of tiering compound over time as data volumes grow.

Invest in a dedicated storage network fabric Sharing the general-purpose data network for storage traffic introduces unpredictable latency and contention that directly impacts bioinformatics pipeline performance. The additional cost of a dedicated fabric is justified by the performance and operational clarity it provides.

Plan for cloud integration as part of the storage architecture Cloud-based archival and analytics are increasingly standard components of genomics data infrastructure. Designing the cloud connectivity and data transfer pipelines as integral parts of the storage architecture avoids the bandwidth bottlenecks and data management inconsistencies that arise when cloud integration is bolted on after the fact.

Document and rehearse critical maintenance procedures The zero-downtime outcome of the power shutdown described in this paper was directly attributable to thorough documentation, stakeholder coordination, and pre-execution review. Any facility that depends on continuous data availability should invest in this level of maintenance planning.

8. Conclusion

This paper has presented a comprehensive account of the storage infrastructure architecture, integration methodology, and operational experience at a national-scale omics data centre supporting one of the largest population genomics programmes in the region. The tiered storage model, zero-downtime commissioning methodology, and cloud-integrated architecture described here provide a replicable framework for similar facilities worldwide.

As genomics programmes continue to scale, processing larger populations, generating multi-omic datasets, and integrating clinical and research workflows, the storage infrastructure that supports them will be a primary determinant of programme success. The experience documented in this paper demonstrates that, with careful design, thorough planning, and disciplined execution, even the most demanding storage infrastructure projects can be completed without compromising the integrity and availability of the critical research operations they support.

Future work should address the emerging challenges of multi-omic data integration, combining genomic, transcriptomic, proteomic, and metabolomic datasets. It should also examine the evolving role of AI-driven storage management systems, and the implications of new sequencing technologies, such as long-read and single-cell sequencing, that generate substantially larger per-sample data volumes and introduce new access patterns.

References

[1] Stephens, Z. D., Lee, S. Y., Faghri, F. et al., Big Data Astronomical or Genomical, PLoS Biology, vol. 13, no. 7, e1002195, 2015.

[2] National Healthcare Technology Organisation, Omics Data Centre Operational Overview, Annual Report, 2023.

[3] Telecommunications Industry Association, TIA-942-B Telecommunications Infrastructure Standard for Data Centers, TIA, 2017.

[4] Uptime Institute, Tier Classification System for Data Centre Infrastructure, Uptime Institute Professional Services, 2023.

[5] Leading Cloud Platform Provider, ExpressRoute and Dedicated Private Circuit: Documentation and Deployment Best Practices, Cloud Platform Technical Documentation, 2024.

[6] National Institute of Standards and Technology, NIST SP 800-53 Revision 5 Security and Privacy Controls for Information Systems and Organizations, U.S. Department of Commerce, 2020.

[7] Department of Health Abu Dhabi, Abu Dhabi Healthcare Information and Cyber Security Standard Version 2.0, Abu Dhabi, 2022.

[8] European Bioinformatics Institute, EMBL-EBI Data Infrastructure Annual Report, EMBL, 2023.

[9] Broad Institute, Genomics Data Infrastructure at Scale, bioRxiv, 2023.

[10] UK Biobank, Data Management Framework and Infrastructure, UK Biobank Technical Reports, 2022.

[11] Muir, P. et al., The Real Cost of Sequencing Scaling Computation to Keep Pace with Data Generation, Genome Biology, vol. 17, no. 1, pp. 53, 2016.

[12] Langmead, B. and Nellore, A., Cloud Computing for Genomic Data Analysis and Collaboration, Nature Reviews Genetics, vol. 19, pp. 208-219, 2018.

[13] Marx, V., The Big Challenges of Big Data, Nature, vol. 498, pp. 255-260, 2013.

[14] Schatz, M. C. and Langmead, B., The DNA Data Deluge, IEEE Spectrum, vol. 50, no. 7, pp. 28-33, 2013.

[15] IEEE Computer Society, Recommended Practice for Data Centre Design and Operations, IEEE, 2022.