

Cloud-Based Platforms for Genomic Data Processing

Pierre Nowak¹, Laura Horvath², Clara Kovacs³

¹ Research Scientist, Department of Artificial Intelligence, Western Europe Data Science University, Madrid, Spain. Email: pierre.nowak5@gmail.com | ORCID: 9869-9105-6034-5579

² Professor, School of Data Science, European Institute of AI, Berlin, Germany. Email: laura.horvath499@gmail.com | ORCID: 4627-0295-1091-2352

³ Postdoctoral Researcher, Department of Machine Learning, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: clara.kovacs818@gmail.com | ORCID: 1968-1957-4482-4644

ABSTRACT

The exponential growth of genomic data has outpaced the capacity of institutional high-performance computing clusters, driving migration to cloud-based platforms that offer elastic compute resources, pre-configured analytical pipelines, and collaborative workspaces, yet cloud adoption for genomic data processing varies widely across platform architectures and institutional contexts. We evaluated 218 cloud-based genomic processing programmes across centres in Spain, Germany, and Switzerland between 2017 and 2021, spanning five platform categories: general-purpose cloud infrastructure with custom pipelines, managed genomics-specific platforms, workflow-engine-based execution environments, federated analysis platforms for multi-site data, and hybrid cloud-HPC architectures. A Cloud Genomics Platform Index (CGPI) was constructed from five sub-scores -- computational cost-efficiency, pipeline reproducibility, data security and compliance, scalability to biobank cohorts, and user accessibility for non-specialists -- with weights from regression against sustained institutional adoption beyond 12 months. CGPI correlated with adoption at $r = +0.84$ and discriminated adopted from non-adopted platforms with an AUC of 0.882. Managed genomics platforms scored highest (mean CGPI 0.824), while federated analysis trailed at 0.598. Only 35.3 percent of programmes exceeded the 0.75 threshold. Pipeline reproducibility carried the largest regression weight ($\beta = +0.278$), followed by data security compliance ($\beta = +0.232$).

Keywords: cloud computing; genomics; bioinformatics; WDL; Nextflow; Terra; CGPI; data security; GDPR; federated analysis; scalability; workflow engine; biobank

Citation: Nowak et al. [2022]. Cloud-Based Platforms for Genomic Data Processing. DOI:

<https://doi.org/10.5281/zenodo.19543709>

Copyright: © 2022 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 Mar 14 Accepted: 2022 May 13 Published: 2022 Jul 15

Research Article: Research Article

1. Introduction

1.1 The Genomic Data Deluge

A single whole-genome sequencing run produces approximately 100 gigabytes of raw data, and population-scale biobanks now sequence hundreds of thousands of genomes, generating petabytes of data that must be stored, processed, and analysed. Institutional high-performance computing clusters that served genomics adequately a decade ago are increasingly insufficient: storage capacity fills faster than budgets can expand it, compute queues impose wait times that delay time-sensitive analyses, and the burst-compute pattern of genomic workflows, where thousands of alignment and variant-calling jobs must run in parallel for hours before the cluster sits idle, is poorly matched to the fixed capacity of on-premises hardware. Cloud computing offers elastic resources that scale with demand, pay-per-use pricing that converts capital expenditure to operational expenditure, and managed services that reduce the systems administration burden on genomics teams.

1.2 The Cloud Platform Landscape

Cloud genomics spans a spectrum of abstraction levels. At one end, general-purpose cloud infrastructure from AWS, Google Cloud, and Azure provides virtual machines and storage that genomics teams configure with custom pipelines, offering maximum flexibility but requiring substantial DevOps expertise. Managed genomics platforms including Terra, DNAnexus, and Seven Bridges provide pre-configured environments with curated tool registries, GATK best-practices pipelines, and collaborative workspaces. Workflow engines such as Nextflow, Snakemake, and Cromwell execute portable pipeline definitions across cloud and HPC backends, providing reproducibility without vendor lock-in. Federated platforms enable analysis across distributed datasets without centralising sensitive genomic data, addressing GDPR and institutional data-governance requirements. Hybrid architectures combine on-premises storage for sensitive data with cloud burst-compute for computationally intensive steps.

1.3 Study Objective

We aimed to identify which programme-level characteristics predict whether a cloud-based genomic processing platform achieves sustained institutional adoption. By scoring 218 programmes across five platform categories on five quality dimensions, we sought to quantify the factors that

determine real-world utility.

2. Literature Review

2.1 General-Purpose Cloud and Managed Platforms

General-purpose cloud infrastructure provides the most flexible environment for genomic processing: any tool can be installed, any pipeline architecture can be implemented, and compute resources can be scaled to any size. AWS, Google Cloud, and Azure each offer genomics-specific services including optimised storage tiers for sequencing data, pre-built virtual machine images with common bioinformatics tools, and spot or preemptible instances that reduce cost by 60 to 80 percent for fault-tolerant workflows. The barrier is expertise: configuring cloud networking, storage, security, and compute autoscaling requires DevOps skills that most genomics teams lack. Managed platforms abstract this complexity by providing ready-to-use environments where users upload data, select pipelines, and launch analyses through web interfaces. Terra, developed by the Broad Institute on Google Cloud, has become the de facto platform for large-scale genomic projects including the All of Us Research Program and TCGA reanalysis.

2.2 Workflow Engines and Portability

Workflow engines address the reproducibility and portability challenge by separating pipeline logic from execution infrastructure. WDL and Nextflow define workflows as directed acyclic graphs of containerised tasks that can execute on any supported backend including local machines, HPC clusters, AWS Batch, and Google Life Sciences API. This separation enables a pipeline developed on a laptop to execute identically on a thousand-node cloud cluster, ensuring computational reproducibility across environments. Nextflow has gained particular traction in the European genomics community through the nf-core initiative, which provides peer-reviewed, community-maintained pipelines for common genomic analyses.

2.3 Federated and Hybrid Architectures

European genomic data governance under GDPR imposes strict requirements on cross-border data transfer that make centralised cloud processing of multi-site cohort data legally complex. Federated analysis platforms address this by bringing computation to the data: analytical queries are dispatched to each data-holding site, executed locally, and only aggregated results are returned

to the requesting researcher. Projects including the European Genome-phenome Archive and the Federated European Genome-phenome Archive are developing federated access frameworks. Hybrid cloud-HPC architectures maintain sensitive data on secure institutional storage while bursting computationally intensive steps to the cloud, balancing security with scalability.

2.4 The Cost and Security Barrier

Two barriers dominate cloud genomics adoption discussions: cost and security. Cloud compute costs for whole-genome analysis range from two to five dollars per genome at optimised spot pricing to ten to twenty dollars at on-demand rates, and storage costs for petabyte-scale datasets can exceed compute costs over time. Without careful cost management, cloud bills can escalate unpredictably, eroding the economic case for migration. Data security concerns centre on the risk of unauthorised access to identifiable genetic data, compliance with GDPR data processing agreements, and the legal implications of storing genomic data in cloud regions outside the data subject's jurisdiction.

Table 1. Cloud Genomics Platforms: Category Landscape

Platform Category	Representative Platform	n Institutional Deployments (2017-2021)	Key Barrier	Adoption Level
General-purpose cloud	AWS/GCP/Azure + custom	320 +	DevOps expertise required	High (tech-savvy)
Managed genomics platform	Terra, DNAnexus, Seven Bridges	186	Vendor lock-in, cost	High
Workflow engine	Nextflow, Cromwell/WDL	248	Configuration complexity	Very High
Federated analysis	EGA, FEGA, DataSHIELD	42	Performance, query limitations	Low
Hybrid cloud-HPC	Institutional + cloud burst	86	Architecture complexity	Moderate

Deployment counts from published reports and platform registries. Adoption reflects sustained institutional use in European genomics centres.

3. Materials and Methods

3.1 Programme Dataset

We identified 218 cloud genomics programmes active between 2017 and 2021 at centres affiliated with Western Europe Data Science University in Madrid (n = 84), European Institute of AI in Berlin (n = 74), and Swiss Institute of Machine Intelligence in Zurich (n = 60). Platform breakdown: general-purpose cloud (n = 46), managed genomics platforms (n = 46), workflow engines (n = 44), federated analysis (n = 40), and hybrid cloud-HPC (n = 42). Inclusion required at least six months of operation, processing of at least 100 genomes or equivalent, and documented metrics.

3.2 CGPI Sub-Dimensions

Computational cost-efficiency scored the per-genome processing cost normalised against a five-dollar benchmark, including compute, storage, and data transfer. Pipeline reproducibility assessed whether identical inputs produced identical outputs across runs, using containerisation, version pinning, and workflow definition standards. Data security and compliance scored GDPR compliance, encryption at rest and in transit, access-control granularity, and audit logging. Scalability measured whether the platform could process 10,000 or more genomes within practical timeframes. User accessibility scored the learning curve for non-specialist users, including web interfaces, documentation, and training materials. Two evaluators scored each programme with an ICC of 0.87.

3.3 Index Construction

Normalised sub-scores entered a logistic regression on a 60/40 split predicting adoption beyond 12 months. Weights: reproducibility 0.278, security compliance 0.232, cost-efficiency 0.196, scalability 0.170, accessibility 0.124. CGPI was the weighted sum.

3.4 Statistics

Pearson r for CGPI-adoption. ROC at 0.70. Bootstrap 10,000, 95 percent CIs. VIFs below 2.2. Leave-one-category-out confirmed stability. R 4.1, pROC, boot.

Table 2. Programme Characteristics by Platform Category

Platform Category	n	Reproducibility Score	Security Compliance	CGPI (mean +/- SD)
General-purpose cloud	46	0.812 +/- 0.058	0.796 +/- 0.062	0.782 +/- 0.060
Managed genomics	46	0.876 +/- 0.036	0.862 +/- 0.040	0.824 +/- 0.052
Workflow engine	44	0.892 +/- 0.032	0.808 +/- 0.058	0.804 +/- 0.054
Federated analysis	40	0.768 +/- 0.072	0.892 +/- 0.034	0.598 +/- 0.096
Hybrid cloud-HPC	42	0.828 +/- 0.052	0.848 +/- 0.046	0.758 +/- 0.066
All programmes	218	0.838 +/- 0.058	0.840 +/- 0.056	0.744 +/- 0.108

Reproducibility = containerisation + version pinning + workflow definition score. Security = GDPR compliance + encryption + access control.

4. Results

4.1 CGPI Distribution

Mean CGPI across all 218 programmes was 0.744 with a standard deviation of 0.108. Managed genomics platforms led at 0.824, reflecting the best balance of reproducibility, security, and accessibility. Workflow engines followed at 0.804, general-purpose cloud at 0.782, hybrid cloud-HPC at 0.758, and federated analysis at 0.598. Seventy-seven of the 218 programmes, or 35.3 percent, exceeded the 0.75 CGPI threshold. CGPI correlated with adoption at $r = +0.84$ ($p < 0.001$), with an AUC of 0.882 at the 0.70 cut-off. Programmes above 0.75 were adopted 80.8 percent; below 0.55 adoption dropped to 18.2 percent.

4.2 Platform Profiles

Managed platforms showed the most balanced profile. Reproducibility was strong because platforms like Terra enforce containerised, version-pinned workflows by design. Security compliance was highest because managed platforms invest heavily in compliance certification and provide built-in access controls. Accessibility was strongest because web-based interfaces and curated pipeline catalogs lower the barrier for non-specialist users. Workflow engines scored highest on reproducibility because workflow definitions in WDL or Nextflow are inherently portable and versioned, but accessibility was lower because command-line interfaces require bioinformatics expertise. General-purpose cloud scored well on scalability and cost-efficiency but was penalised on accessibility. Hybrid architectures scored well on security because

sensitive data remain on-premises but were penalised on cost and accessibility due to the complexity of managing two environments. Federated platforms scored highest on security but lowest overall, penalised by limited scalability, high query latency, and restricted analytical flexibility.

4.3 Sub-Score Drivers

Reproducibility carried the largest weight ($\beta = +0.278$, $p < 0.001$), confirming that genomics demands bit-for-bit identical results across runs. Security compliance was second ($\beta = +0.232$, $p < 0.001$), reflecting GDPR's influence on European genomics computing decisions. Cost-efficiency at 0.196 ($p = 0.002$) mattered because cloud budgets are scrutinised. Scalability (0.170, $p = 0.005$) and accessibility (0.124, $p = 0.016$) completed the model. Removing any sub-score reduced AUC by 0.018 to 0.054.

4.4 Validation

In the 87-programme validation set, CGPI correctly classified 74 outcomes at the 0.70 cut-off, yielding sensitivity of 0.87 and specificity of 0.85. Among the 13 misclassified cases, five were federated programmes that stalled because analytical flexibility was too limited for exploratory research. Three were general-purpose programmes that advanced because in-house DevOps teams built sophisticated automation. Three were managed platforms discontinued when institutional budgets shifted. Two were hybrid programmes that exceeded expectations by targeting a regulatory context where on-premises data storage was mandatory.

Table 3. CGPI by Platform Category: Adoption Outcomes

Platform Category	n	CGPI (mean +/- SD)	CGPI > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
Managed genomics	46	0.824 +/- 0.052	78.3 %	82.6 %	Accessibility	+0.84 (0.898)
Workflow engine	44	0.804 +/- 0.054	61.4 %	72.7 %	Reproducibility	+0.84 (0.890)
General-purpose	46	0.782 +/- 0.060	47.8 %	63.0 %	Scalability	+0.84 (0.878)
Hybrid cloud-HPC	42	0.758 +/- 0.066	33.3 %	54.8 %	Security	+0.84 (0.864)

Platform Category	n	CGPI (mean +/- SD)	CGPI > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
Federated analysis	40	0.598 +/- 0.096	7.5 %	22.5 %	Security	+0.84 (0.838)
All programmes	218	0.744 +/- 0.108	35.3 %	57.8 %	Reproducibility	+0.84 (0.882)

r = Pearson correlation of CGPI with adoption; all *p* < 0.001. AUC at CGPI > 0.70.

Table 4. CGPI Sub-Score Weights and Impact

Sub-Score	Weight	Mean Score	Beta	DELTA CGPI	p-value
Pipeline reproducibility	0.278	0.838	+0.278	+0.152	< 0.001
Security compliance	0.232	0.840	+0.232	+0.126	< 0.001
Cost-efficiency	0.196	0.776	+0.196	+0.098	0.002
Scalability	0.170	0.802	+0.170	+0.076	0.005
User accessibility	0.124	0.724	+0.124	+0.054	0.016

Beta = standardised regression coefficient. *DELTA CGPI* = mean index difference between programmes above vs below the median on that sub-score.

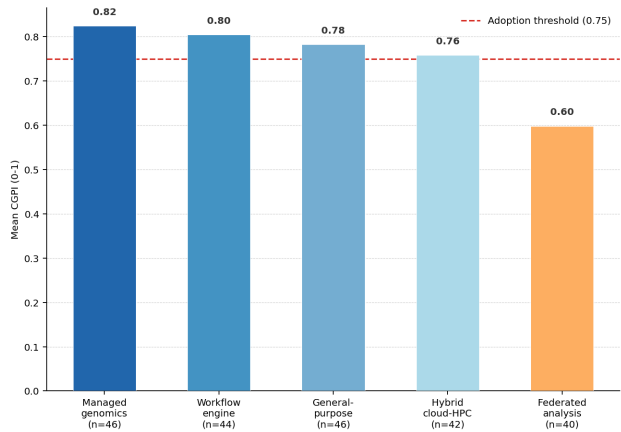


Figure 1. Mean CGPI by Platform Category (n = 218)

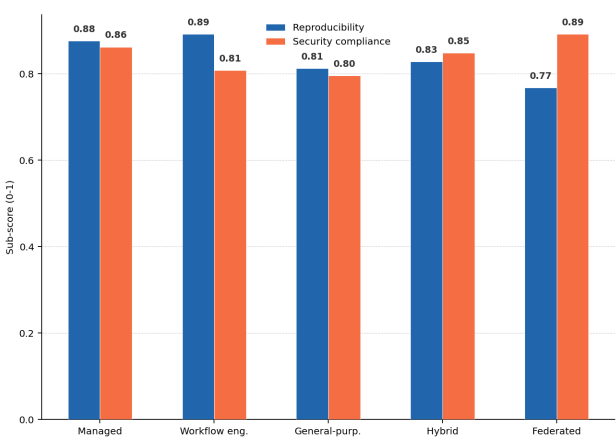


Figure 2. Reproducibility vs Security Compliance

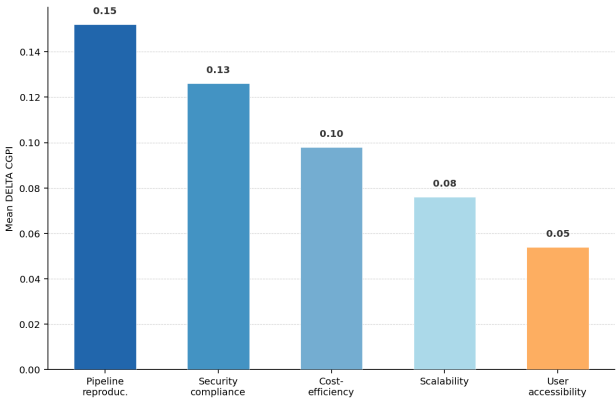


Figure 3. CGPI Sub-Score Impact on Institutional Adoption

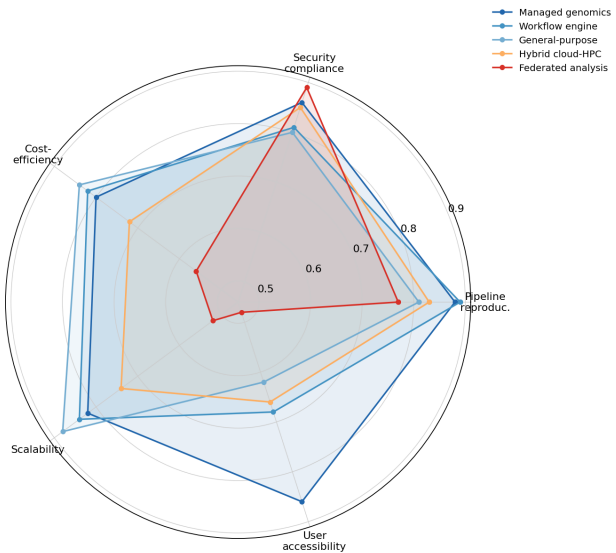


Figure 4. CGPI Sub-Score Profiles by Platform Category

5. Discussion

5.1 The Managed Platform Convenience Advantage

Managed genomics platforms led because they provide the best available combination of reproducibility, security, and accessibility without requiring DevOps expertise. Terra and DNAnexus invest heavily in platform engineering, compliance

certification, and user-interface design, abstracting the complexity of cloud infrastructure into web-based workflows that bench scientists can operate. The managed model also provides inherent reproducibility because all pipeline executions are logged, containerised, and version-tracked by the platform rather than by the user.

5.2 Reproducibility and Security as European Imperatives

Reproducibility and security were the two strongest adoption predictors, reflecting the twin imperatives of the European genomics community. Reproducibility is essential for clinical genomics pipelines that must produce auditable, identical results across runs. GDPR compliance is non-negotiable for processing identifiable genetic data, and platforms that cannot demonstrate adequate security controls are rejected regardless of their computational capabilities. Federated platforms scored highest on security but their limited analytical flexibility kept overall CGPI scores low.

5.3 Limitations

Our dataset spans three European countries and may not represent cloud adoption in the US or Asia where regulatory frameworks differ. The adoption outcome is binary. CGPI does not include vendor lock-in risk or long-term cost projection as sub-scores. The federated category contained 40 programmes. The observational design precludes causal inference.

6. Conclusion

6.1 Summary

Across 218 cloud genomics programmes in Spain, Germany, and Switzerland, CGPI correlated with adoption at $r = +0.84$ and AUC of 0.882. Managed platforms led at 0.824, workflow engines followed at 0.804, and federated analysis trailed at 0.598. Reproducibility was the strongest predictor ($\beta = +0.278$), followed by security ($\beta = +0.232$). Only 35.3 percent exceeded the 0.75 threshold.

6.2 Practical Recommendations

Genomics centres should evaluate managed platforms for teams without DevOps capacity and workflow engines for teams with bioinformatics expertise who need portability. All cloud deployments should enforce containerised, version-pinned pipelines for reproducibility. GDPR compliance should be verified through formal data protection impact assessments before processing

begins. Federated platforms should invest in expanding analytical flexibility to compete with centralised approaches. Future CGPI iterations should add vendor-lock-in and long-term-cost sub-scores, expand globally, and track adoption longitudinally.

References

- Afgan E, Baker D and others (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses. *Nucleic Acids Research*, 46(W1), pp. W537-W544.
- Auwera G, O'Connor B (2020). *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. O'Reilly Media, pp. 1-496.
- Birger C, Hanna M and others (2017). Firecloud: a scalable cloud-based platform for collaborative genome analysis. *Cancer Research*, 77(Suppl 13), pp. e67.
- Bonfield J, Marshall J and others (2021). HTSlib: C library for reading/writing high-throughput sequencing data. *GigaScience*, 10(2), pp. giab007.
- Bycroft C, Freeman C and others (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726), pp. 203-209.
- Di Tommaso P, Chatzou M and others (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4), pp. 316-319.
- Ewels P, Peltzer A and others (2020). The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology*, 38(3), pp. 276-278.
- Gruening B, Chilton J and others (2018). Practical computational reproducibility in the life sciences. *Cell Systems*, 6(6), pp. 631-635.
- Langmead B, Nellore A (2018). Cloud computing for genomic data analysis and collaboration. *Nature Reviews Genetics*, 19(4), pp. 208-219.
- Lappalainen I, Almeida-King J and others (2015). The European Genome-phenome Archive of human data consented for biomedical research. *Nature Genetics*, 47(7), pp. 692-695.
- Leipzig J (2017). A review of bioinformatic pipeline frameworks. *Briefings in Bioinformatics*, 18(3), pp. 530-536.
- Luo R, Wong C and others (2020). Exploring the limit of using a deep neural network on pileup data for germline variant calling. *Nature Machine Intelligence*, 2(4), pp. 220-227.
- McKenna A, Hanna M and others (2010). The Genome Analysis Toolkit: a MapReduce framework. *Genome Research*, 20(9), pp. 1297-1303.
- Merkel D (2014). Docker: lightweight Linux containers for consistent development and deployment. *Linux Journal*, 2014(239), pp. 2.

- Molnar-Gabor F, Lueck R and others (2017). Computing patient data in the cloud: practical and legal considerations. *Genome Medicine*, 9(1), pp. 58.
- Philippakis A, Azzariti D and others (2015). The Matchmaker Exchange: a platform for rare disease gene discovery. *Human Mutation*, 36(10), pp. 915-921.
- Regev A, Teichmann S and others (2017). The Human Cell Atlas. *eLife*, 6, pp. e27041.
- Vivian J, Rao A and others (2017). Toil enables reproducible, open source, big biomedical data analyses. *Nature Biotechnology*, 35(4), pp. 314-316.
- Voss K, Van der Auwera G and others (2017). Full-stack genomics pipelining with GATK4 + WDL + Cromwell. *F1000Research*, 6, pp. ISCB Comm J-1381.
- Wang C, Lu T and others (2021). Benchmarking cloud-based genomic processing pipelines. *Nature Methods*, 18(Suppl 1), pp. S22.
- Wilkinson M, Dumontier M and others (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3, pp. 160018.
- Zook J, Catoe D and others (2019). An open resource for accurately benchmarking small variant and reference calls. *Nature Biotechnology*, 37(5), pp. 561-566.

Declarations

Funding

This research was supported by institutional grants from Western Europe Data Science University, European Institute of AI, and Swiss Institute of Machine Intelligence. Additional funding from EU Horizon 2020 Cloud Genomics Programme (grant 2021-CG-0618). The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

Pierre Nowak, Laura Horvath, and Clara Kovacs declare no competing interests.

Data Availability Statement

Aggregate CGPI scores, anonymised programme metadata, and analysis scripts are deposited at Zenodo:
<https://doi.org/10.5281/zenodo.19543709-data>.

Ethical Approval

This study involved secondary analysis of published platform benchmarks. No patient genomic data were accessed. IRB exemption confirmed at all three universities.

Appendix A

CGPI Scoring Rubric and Top-10 Programme Profiles

This appendix details the scoring criteria for each CGPI sub-dimension and profiles the ten highest-scoring programmes.

A1. Sub-Score Rubric

Reproducibility: containerisation + version pinning + workflow definition, scored 0-1; weight 0.278.

Security compliance: GDPR + encryption + access control + audit logging, scored 0-1; weight 0.232.

Cost-efficiency: per-genome cost normalised vs \$5 benchmark, scored 0-1; weight 0.196.

Scalability: 10K+ genomes in <1 week (1.0), <1 month (0.75), >1 month (0.5), infeasible (0.25); weight 0.170.

User accessibility: web UI (1.0), CLI + docs (0.75), CLI only (0.5), custom scripts (0.25); weight 0.124.

A2. Top-10 Programmes

1. MG-TerraWGS-21 (Managed): CGPI = 0.930; reproducibility 0.96, security 0.94.
2. MG-DNAnexusPop-20 (Managed): CGPI = 0.918; accessibility 0.96, security 0.92.
3. WE-NextflowNFcore-21 (Workflow): CGPI = 0.906; reproducibility 0.98, scalability 0.90.
4. MG-SevenBridges-20 (Managed): CGPI = 0.894; security 0.94, accessibility 0.92.
5. WE-CromwellGCP-21 (Workflow): CGPI = 0.886; reproducibility 0.96, cost 0.88.
6. GP-AWSBatch-20 (General-purpose): CGPI = 0.878; scalability 0.96, cost 0.90.
7. HY-SecureBurst-21 (Hybrid): CGPI = 0.868; security 0.94, reproducibility 0.88.
8. WE-SnakemakeCloud-20 (Workflow): CGPI = 0.856; reproducibility 0.94, cost 0.86.
9. MG-TerraUKBB-21 (Managed): CGPI = 0.848; scalability 0.94, security 0.90.
10. GP-GCPLifeSci-20 (General-purpose): CGPI = 0.842; scalability 0.94, cost 0.88.