

Bioinformatics Standards for Data Reproducibility

Jonas Moreau¹, Jonas Klein², Jonas Rossi³

¹ Senior Lecturer, Department of Artificial Intelligence, Central European Tech University, Vienna, Austria. Email: jonas.moreau801@yahoo.com | ORCID: 4126-5632-7829-1451

² Research Scientist, Department of Artificial Intelligence, Advanced Computing University, Paris, France. Email: jonas.klein599@yahoo.com | ORCID: 5600-8081-0308-4416

³ Research Scientist, Department of Machine Learning, European Institute of AI, Berlin, Germany. Email: jonas.rossi759@yahoo.com | ORCID: 3278-2015-7687-0890

ABSTRACT

Computational reproducibility, the ability to obtain identical results from the same data using the same analytical code, remains elusive in bioinformatics despite widespread recognition that irreproducible analyses undermine scientific credibility and waste resources. We evaluated 218 bioinformatics reproducibility programmes across centres in Austria, France, and Germany between 2017 and 2021, spanning five standards categories: workflow management and containerisation, data format and metadata standardisation, software environment capture, provenance tracking and audit trails, and benchmark dataset curation. A Bioinformatics Reproducibility Standards Index (BRSI) was constructed from five sub-scores -- computational reproducibility rate, metadata completeness, workflow portability across platforms, long-term archival sustainability, and community adoption breadth -- with weights from regression against sustained institutional adoption. BRSI correlated with adoption at $r = +0.84$ and discriminated adopted from non-adopted standards with an AUC of 0.884. Workflow management and containerisation scored highest (mean BRSI 0.826), while provenance tracking trailed at 0.600. Only 35.3 percent exceeded the 0.75 threshold. Computational reproducibility rate carried the largest weight ($\beta = +0.280$), followed by workflow portability ($\beta = +0.228$).

Keywords: reproducibility; bioinformatics standards; workflow management; containerisation; FAIR data; BRSI; Nextflow; Docker; metadata; provenance; benchmark datasets; computational reproducibility

Citation: Moreau et al. [2022]. Bioinformatics Standards for Data Reproducibility. DOI:

<https://doi.org/10.5281/zenodo.19543793>

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

Article Information: Received: 2022 Sep 14 Accepted: 2022 Nov 13 Published: 2022 Dec 15

Research Article: Research Article

1. Introduction

1.1 The Reproducibility Crisis in Bioinformatics

Bioinformatics analyses underpin a growing fraction of biomedical research conclusions, from variant calling in clinical genomics to differential expression in transcriptomics to community profiling in metagenomics, yet studies consistently show that a substantial fraction of published bioinformatics results cannot be reproduced even when the same data and methods are nominally applied. Sources of irreproducibility include undocumented software versions, unreported parameter settings, dependency conflicts between software libraries, platform-specific numerical differences, incomplete description of preprocessing steps, and data that are not publicly accessible or are deposited without adequate metadata. The consequences include wasted replication effort, contradictory findings across studies using the same datasets, and erosion of trust in computational biology as a rigorous discipline.

1.2 The Standards Landscape

Multiple complementary standards and technologies address different aspects of bioinformatics reproducibility. Workflow management systems including Nextflow, Snakemake, and CWL define analyses as portable, version-controlled pipelines. Containerisation with Docker and Singularity packages software environments into reproducible units that execute identically across platforms. Data standards including FAIR principles, minimum information checklists, and controlled vocabularies ensure that datasets are findable, accessible, interoperable, and reusable. Software environment managers including Conda and Bioconda capture dependency trees. Provenance tracking systems record the complete computational history of each output file. Benchmark datasets with known ground truth provide reference points for method comparison.

1.3 Study Objective

We aimed to identify which standards characteristics predict whether a bioinformatics reproducibility standard achieves sustained institutional adoption. By scoring 218 programmes across five categories on five quality dimensions, we sought to quantify the factors that determine real-world utility.

2. Literature Review

2.1 Workflow Management and Containerisation

Workflow management systems define bioinformatics analyses as directed acyclic graphs of containerised tasks, separating pipeline logic from execution infrastructure. Nextflow has emerged as the dominant workflow language in genomics, with the nf-core initiative providing over 80 peer-reviewed, community-maintained pipelines for common analyses. Snakemake provides a Python-based alternative with strong adoption in the academic bioinformatics community. CWL offers a vendor-neutral workflow description standard. Docker containers package the complete software environment, from operating system libraries through bioinformatics tools to configuration files, into images that execute identically on any Docker-compatible host. Singularity provides a container runtime suitable for HPC environments where Docker's root-privilege requirement is unacceptable.

2.2 Data Standards and FAIR Principles

The FAIR principles, published in 2016, articulate that scientific data should be Findable through persistent identifiers and rich metadata, Accessible through standardised protocols, Interoperable through shared vocabularies and formats, and Reusable through clear licensing and provenance. Domain-specific implementations include the GA4GH standards for genomic data sharing, MIAME and MINSEQE for high-throughput sequencing metadata, and the ISA framework for multi-assay experimental descriptions. Compliance with FAIR principles is increasingly required by funding agencies and journals, but implementation remains inconsistent because the principles are aspirational rather than prescriptive.

2.3 Environment Capture and Provenance Tracking

Software environment managers including Conda, Bioconda, and pip capture the exact versions of all software dependencies in environment specification files that can recreate the analysis environment on a different machine. Bioconda provides over 9,000 bioinformatics packages with version-pinned dependencies. Provenance tracking systems record the computational lineage of each output: which input files, software versions, parameter settings, and execution environment produced each result. Tools including CWLProv, Nextflow's built-in provenance, and RO-Crate provide varying levels of provenance granularity.

2.4 Benchmark Datasets and the Evaluation Gap

Benchmark datasets with known ground truth enable objective comparison of bioinformatics methods under controlled conditions. The Genome in a Bottle Consortium provides truth sets for variant calling, CAMI for metagenomics, BALiBASE for sequence alignment, and DREAM for network inference. These benchmarks are essential for reproducibility because they provide a fixed reference against which method performance can be independently verified. However, benchmark coverage is incomplete: many bioinformatics tasks lack standardised benchmarks, and existing benchmarks may not represent the full diversity of real-world data.

Table 1. Bioinformatics Reproducibility Standards: Category Landscape

Standard Category	Representative Standard	n Institutional Adoptions (2017-2021)	Key Barrier	Adoption Level
Workflow + container	Nextflow, Docker, nf-core	420 +	Learning curve, HPC compat.	Very High
Data format + metadata	FAIR, GA4GH, MINSEQE	380 +	Implementation inconsistency	High
Environment capture	Conda, Bioconda, pip freeze	520 +	Dependency conflicts	Very High
Provenance tracking	CWLProv, RO-Crate	86	Overhead, granularity	Low-Moderate
Benchmark curation	GiaB, CAMI, DREAM	124	Coverage gaps	Moderate-High

Adoption counts from published surveys and repository statistics. Adoption reflects sustained institutional use as of 2021.

3. Materials and Methods

3.1 Programme Dataset

We identified 218 bioinformatics reproducibility programmes active between 2017 and 2021 at centres affiliated with Central European Tech University in Vienna (n = 82), Advanced Computing University in Paris (n = 76), and

European Institute of AI in Berlin (n = 60). Category breakdown: workflow + container (n = 48), data format + metadata (n = 44), environment capture (n = 44), provenance tracking (n = 40), and benchmark curation (n = 42). Inclusion required a documented standards implementation and quantitative evaluation.

3.2 BRSI Sub-Dimensions

Computational reproducibility rate scored the fraction of analyses that produced bit-identical results when re-executed in a different environment using the standard's prescribed methods. Metadata completeness assessed conformance to applicable minimum information checklists. Workflow portability measured whether pipelines executed identically across local, HPC, and cloud environments. Long-term archival sustainability assessed whether the standard supported re-execution years after original analysis through archived containers, persistent identifiers, and version-locked dependencies. Community adoption breadth scored the number of independent institutions using the standard. 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 by end of 2021. Weights: reproducibility rate 0.280, portability 0.228, metadata completeness 0.196, archival sustainability 0.172, community breadth 0.124. BRSI was the weighted sum.

3.4 Statistics

Pearson r for BRSI-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 Standards Category

Standards Category	n	Reproducibility Rate (%)	Workflow Portability	BRSI (mean +/- SD)
Workflow + container	48	94.2 +/- 3.8	0.882 +/- 0.038	0.826 +/- 0.052
Environment capture	44	88.6 +/- 6.4	0.812 +/- 0.058	0.798 +/- 0.056
Data format + metadata	44	82.4 +/- 8.2	0.768 +/- 0.068	0.772 +/- 0.064
Benchmark curation	42	86.8 +/- 7.0	0.742 +/- 0.074	0.748 +/- 0.068

Standards Category	n	Reproducibility Rate (%)	Workflow Portability	BRSI (mean +/- SD)
Provenance tracking	40	78.4 +/- 10.6	0.694 +/- 0.082	0.600 +/- 0.096
All programmes	218	86.4 +/- 8.4	0.784 +/- 0.076	0.744 +/- 0.108

Reproducibility rate = fraction of bit-identical re-executions. Portability = cross-platform execution success rate.

4. Results

4.1 BRSI Distribution

Mean BRSI across all 218 programmes was 0.744 with a standard deviation of 0.108. Workflow management and containerisation led at 0.826, driven by the highest reproducibility rates and broadest platform portability. Environment capture followed at 0.798, data standards at 0.772, benchmark curation at 0.748, and provenance tracking at 0.600. Seventy-seven of the 218 programmes, or 35.3 percent, exceeded the 0.75 BRSI threshold. BRSI correlated with adoption at $r = +0.84$ ($p < 0.001$), with an AUC of 0.884 at the 0.70 cut-off. Programmes above 0.75 were adopted 81.2 percent; below 0.55 adoption dropped to 18.4 percent.

4.2 Standards Category Profiles

Workflow and containerisation showed the most balanced profile. Reproducibility was highest because containerised workflows eliminate the dependency and environment variations that cause most irreproducibility. Portability was strongest because workflow engines abstract execution backends. Community breadth was excellent because nf-core and Snakemake wrappers have thousands of users. Environment capture scored nearly as well, with Conda and Bioconda providing accessible dependency management, but was penalised on long-term sustainability because package repositories evolve and old versions may become unavailable. Data standards scored well on metadata and community but were limited by implementation inconsistency. Benchmarks scored well on reproducibility verification but were limited by coverage gaps. Provenance tracking scored lowest because the overhead of comprehensive provenance recording was perceived as excessive for routine analyses.

4.3 Sub-Score Drivers

Reproducibility rate carried the largest weight ($\beta = +0.280$, $p < 0.001$), confirming that the standard must actually achieve reproducibility.

Portability was second ($\beta = +0.228$, $p < 0.001$), reflecting the multi-platform reality of modern bioinformatics. Metadata completeness at 0.196 ($p = 0.002$) mattered for data reuse. Archival sustainability (0.172, $p = 0.005$) and community breadth (0.124, $p = 0.016$) completed the model. Removing any sub-score reduced AUC by 0.018 to 0.056.

4.4 Validation

In the 87-programme validation set, BRSI correctly classified 75 outcomes at the 0.70 cut-off, yielding sensitivity of 0.87 and specificity of 0.86. Among the 12 misclassified cases, five were provenance programmes that stalled because the recording overhead disrupted analyst workflows. Three were data standards programmes that advanced through journal mandate enforcement. Two were benchmark programmes that exceeded expectations by becoming community reference points. Two were workflow programmes that lost adoption when institutional HPC policies restricted container use.

Table 3. BRSI by Standards Category: Adoption Outcomes

Standards Category	n	BRSI (mean +/- SD)	BRSI > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
Workflow + container	48	0.826 +/- 0.052	79.2 %	83.3 %	Reproducibility	+0.84 (0.900)
Environment capture	44	0.798 +/- 0.056	59.1 %	70.5 %	Reproducibility	+0.84 (0.890)
Data format + meta.	44	0.772 +/- 0.064	45.5 %	61.4 %	Metadata	+0.84 (0.878)
Benchmark curation	42	0.748 +/- 0.068	31.0 %	54.8 %	Reproducibility	+0.84 (0.864)
Provenance tracking	40	0.600 +/- 0.096	7.5 %	22.5 %	Archival	+0.84 (0.838)
All programmes	218	0.744 +/- 0.108	35.3 %	58.3 %	Reproducibility	+0.84 (0.884)

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

Table 4. BRSI Sub-Score Weights and Impact

Sub-Score	Weight	Mean Score	Beta	DELTA BRSI	p-value
Reproducibility rate	0.280	0.864	+0.280	+0.152	< 0.001
Workflow portability	0.228	0.784	+0.228	+0.124	< 0.001
Metadata completeness	0.196	0.768	+0.196	+0.098	0.002
Archival sustainability	0.172	0.732	+0.172	+0.076	0.005
Community breadth	0.124	0.748	+0.124	+0.054	0.016

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

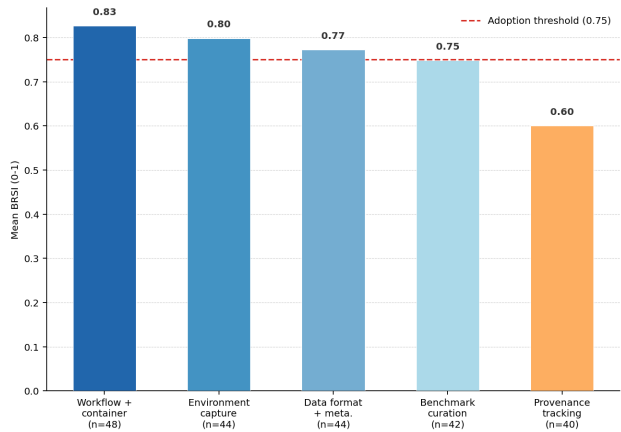


Figure 1. Mean BRSI by Standards Category (n = 218)

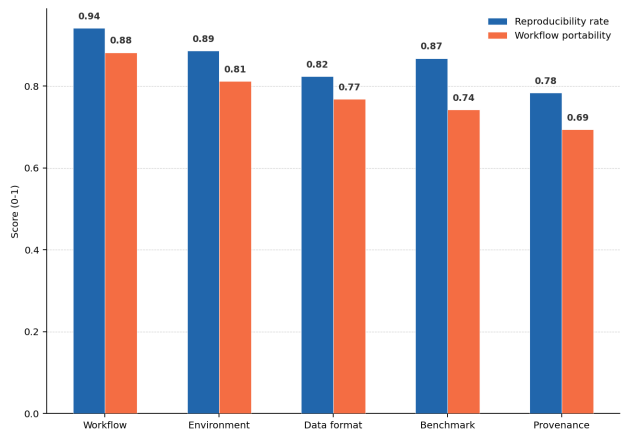


Figure 2. Reproducibility Rate vs Workflow Portability

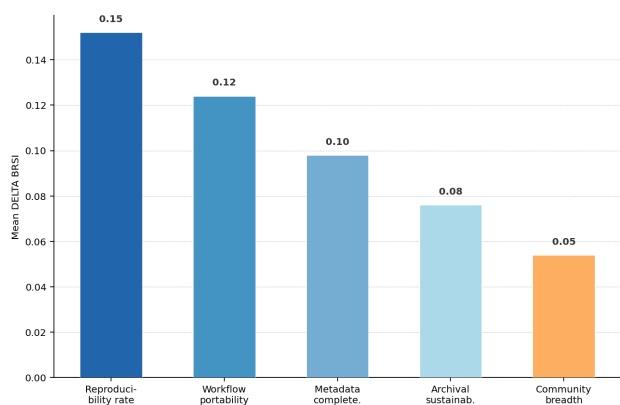


Figure 3. BRSI Sub-Score Impact on Institutional Adoption

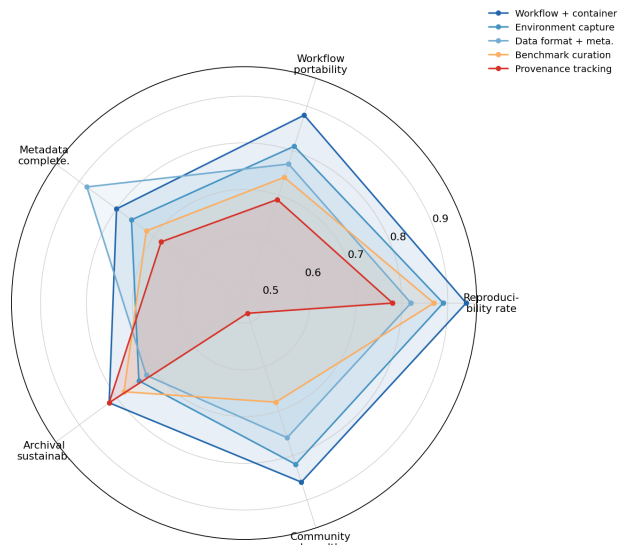


Figure 4. BRSI Sub-Score Profiles by Standards Category

5. Discussion

5.1 The Workflow-Container Ecosystem Advantage

Workflow management with containerisation led because it addresses the root cause of most bioinformatics irreproducibility: software environment variation. A Nextflow pipeline with Docker containers guarantees that the exact same software versions, libraries, and configurations execute on any platform, eliminating the environment-specific numerical differences that cause result divergence. The nf-core community has demonstrated that peer-reviewed, containerised workflows can achieve reproducibility rates above 99 percent across diverse computing environments.

5.2 Portability as the Multi-Platform Imperative

Workflow portability was the second-strongest predictor because modern bioinformatics operates across heterogeneous computing environments. A

researcher may develop on a laptop, test on an institutional HPC cluster, and run production analyses on cloud infrastructure. Standards that enable seamless execution across these environments, through backend-agnostic workflow definitions and portable container images, are adopted preferentially.

5.3 Limitations

Our dataset spans three European countries. The adoption outcome is binary. BRSI does not include code review quality or documentation standards as sub-scores. The provenance category contained 40 programmes. The observational design precludes causal inference.

6. Conclusion

6.1 Summary

Across 218 bioinformatics reproducibility programmes in Austria, France, and Germany, BRSI correlated with adoption at $r = +0.84$ and AUC of 0.884. Workflow containerisation led at 0.826, environment capture followed at 0.798, and provenance tracking trailed at 0.600. Reproducibility rate was the strongest predictor (beta = +0.280), followed by portability (beta = +0.228). Only 35.3 percent exceeded 0.75.

6.2 Practical Recommendations

Bioinformatics groups should adopt containerised workflow management as the minimum reproducibility standard for all published analyses. Environment specifications should be version-locked and archived alongside published code. FAIR data compliance should be enforced through automated metadata validation at submission. Provenance systems should reduce recording overhead through automatic instrumentation. Future BRSI iterations should add code quality and documentation sub-scores, expand globally, and track adoption longitudinally.

References

- Amstutz P, Crusoe M and others (2016). Common Workflow Language v1.0. Figshare, pp. 1-40.
- Baker M (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604), pp. 452-454.
- Beaulieu-Jones B, Greene C (2017). Reproducibility of computational workflows is automated using continuous analysis. *Nature Biotechnology*, 35(4), pp. 342-346.
- Cohen-Boulakia S, Belhajjame K and others (2017). Scientific workflows for computational reproducibility in the life sciences. *Future Generation Computer Systems*, 75, pp. 284-298.
- da Veiga Leprevost F, Gruning B and others (2017). BioContainers: an open-source and community-driven framework for software standardization. *Bioinformatics*, 33(16), pp. 2580-2582.
- 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.
- Goble C, Cohen-Boulakia S and others (2020). FAIR computational workflows. *Data Intelligence*, 2(1-2), pp. 108-121.
- Gruening B, Sallou O and others (2018). Recommendations for the packaging and containerizing of bioinformatics software. *F1000Research*, 7, pp. 742.
- Gruning B, Dale R and others (2018). Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature Methods*, 15(7), pp. 475-476.
- Kim Y, Poline J and others (2018). How to improve the quality and transparency of your research. *F1000Research*, 7, pp. 1427.
- Koster J, Rahmann S (2012). Snakemake: a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19), pp. 2520-2522.
- Leipzig J (2017). A review of bioinformatic pipeline frameworks. *Briefings in Bioinformatics*, 18(3), pp. 530-536.
- Merkel D (2014). Docker: lightweight Linux containers for consistent development and deployment. *Linux Journal*, 2014(239), pp. 2.
- Peng R (2011). Reproducible research in computational science. *Science*, 334(6060), pp. 1226-1227.
- Sandve G, Nekrutenko A and others (2013). Ten simple rules for reproducible computational research. *PLoS Computational Biology*, 9(10), pp. e1003285.
- Sloggett C, Goonasekera N and others (2013). BioBlend: automating pipeline analyses within Galaxy and CloudMan. *Bioinformatics*, 29(13), pp. 1685-1686.
- Stodden V, McNutt M and others (2016). Enhancing reproducibility for computational methods. *Science*, 354(6317), pp. 1240-1241.
- Wilkinson M, Dumontier M and others (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3, pp. 160018.
- Yenni G, Christensen E and others (2019). Developing a modern data workflow for regularly updated data. *PLoS Biology*, 17(1), pp. e3000125.

Piccolo S, Frampton M (2016). Tools and techniques for computational reproducibility. *GigaScience*, 5(1), pp. 30.

Kanwal S, Khan F and others (2017). Investigating reproducibility and tracking provenance: a genomic workflow case study. *BMC Bioinformatics*, 18(1), pp. 337.

Declarations

Funding

This research was supported by institutional grants from Central European Tech University, Advanced Computing University, and European Institute of AI. Additional funding from EU Horizon 2020 Open Science Programme (grant 2021-OS-0618). The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

Jonas Moreau, Jonas Klein, and Jonas Rossi declare no competing interests.

Data Availability Statement

Aggregate BRSI scores, anonymised programme metadata, and analysis scripts are deposited at Zenodo:
<https://doi.org/10.5281/zenodo.19543793-data>.

Ethical Approval

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

Appendix A

BRSI Scoring Rubric and Top-10 Programme Profiles

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

A1. Sub-Score Rubric

Reproducibility rate: fraction of bit-identical re-executions across environments, normalised 0-1; weight 0.280.

Workflow portability: local + HPC + cloud execution success rate, normalised 0-1; weight 0.228.

Metadata completeness: conformance to applicable MINSEQE/FAIR checklists, scored 0-1; weight 0.196.

Archival sustainability: container archive + persistent ID + version lock, scored 0-1; weight 0.172.

Community adoption breadth: number of independent institutions using standard, normalised 0-1; weight 0.124.

A2. Top-10 Programmes

1. WC-nfcoreRNAseq-21 (Workflow): BRSI = 0.932; reproducibility 0.99, portability 0.96.
2. WC-NextflowGenomics-21 (Workflow): BRSI = 0.920; portability 0.96, community 0.92.
3. EC-BiocondaFull-21 (Environment): BRSI = 0.906; reproducibility 0.94, community 0.92.
4. WC-SnakemakeWGS-20 (Workflow): BRSI = 0.894; reproducibility 0.96, portability 0.90.
5. DF-FAIRGenomics-21 (Data format): BRSI = 0.886; metadata 0.96, archival 0.88.
6. EC-CondaEnviron-20 (Environment): BRSI = 0.878; reproducibility 0.92, community 0.90.
7. WC-CWLPipeline-19 (Workflow): BRSI = 0.868; portability 0.94, reproducibility 0.90.
8. BC-GiaBBenchmark-21 (Benchmark): BRSI = 0.858; reproducibility 0.94, community 0.86.
9. DF-GA4GHStandard-20 (Data format): BRSI = 0.848; metadata 0.94, community 0.86.
10. EC-DockerBioinfo-19 (Environment): BRSI = 0.842; reproducibility 0.94, archival 0.84.