

The Case for Data Provenance and Authenticity in Genomics

Building Trustworthy Foundations for Digital Biology

Jonathan L. Jacobs, PhD^{1*}

1. ATCC, 10801 University Blvd., Manassas, VA 20110

* correspondence should be sent to Jonathan L. Jacobs, jjacobs@atcc.org

Abstract

The exponential growth of publicly accessible genomic data over the last two decades has transformed life sciences, yet it has also exposed a critical vulnerability. Weakly enforced requirements for data provenance, structured metadata, and material authentication have degraded the potential of these resources for interoperability and reuse in digital biology. The lack of traceability and verification in genomic data poses escalating risks to scientific reproducibility, biosecurity, and the integrity of AI-driven biological research (AIxBio). Examples from cancer and microbial genomics, infectious disease surveillance, public sequence archives, and emerging AI-enabled biology demonstrate how poor data provenance and metadata quality gaps undermine trust, drive irreproducible results, and create opportunities for data fabrication and misuse. The manuscript further emphasizes that reproducibility alone is insufficient when shared reference data are contaminated, mislabeled, incompletely described, or biologically outdated. Furthermore, the unique role of biological repositories and international culture collections is presented as bridging the physical-to-digital divide and enabling the creation of trusted “digital twins” for biological research. Finally, the proactive preservation of physical reference materials underpinning genomic data and an emphasis on “metadata as infrastructure” is presented as a key ingredient for the future success and sustainability of artificial intelligence and machine learning across the life sciences (i.e., AIxBio). Finally, proactive preservation of physical reference materials and the treatment of “metadata as infrastructure” are presented as key ingredients for the future success and sustainability of artificial intelligence and machine learning across the life sciences.

Introduction

Genomic data have become a cornerstone of modern life sciences research, supporting applications that range from tracking infectious disease outbreaks to developing precision diagnostics, engineering new biotechnologies, and advancing our understanding of human health. Since the creation of the Los Alamos Sequence Data Bank in 1979 (1) and the

National Institutes of Health GenBank database in 1982 (2), the amount of DNA sequence data available to the global research community has roughly doubled every 24 months (3). This rapid expansion has vastly democratized access to genomic information and fueled countless discoveries. However, it has also exposed a critical weakness in our scientific infrastructure: the lack of robust requirements for data provenance and authenticity in genomic databases. As a result, downstream applications that depend on accurate metadata and sample traceability increasingly face significant challenges in interoperability and reproducibility (4–7).

Walter Goad’s 1979 proposal to create the Los Alamos DNA Data Bank acknowledged that the credibility of sequence records could not rest solely on community trust, proposing instead that such a database should include a graded “validation status” tied to objective supporting evidence and independent confirmation (1). In this framework, the provenance of the data, including the methods used to produce it and the source of the original materials, was treated as an artifact equal in importance to the DNA sequence data itself. When the NIH took over management of the Los Alamos data repository and created GenBank in 1982, the NIH ended the practice of assigning a “validation status” for DNA sequence data deposited into the database. Over 40 years later, the primary international sequence repositories (i.e., the International Nucleotide Sequence Database Collaboration [INSDC], which includes GenBank, the European Nucleotide Archive [ENA], and the DNA Data Bank of Japan [DDBJ]) no longer verify the accuracy or source of the sequences they host. Instead, they operate under a shared model in which the responsibility for record quality and accuracy rests primarily with the submitting author, rather than the database itself (8). Accordingly, INSDC databases function primarily as data archives, not curated repositories of validated reference data. Arguably, given the scale of these databases and the pace at which data are deposited into them, the role of INSDC member databases should not be to “police” the veracity of all data being submitted for distribution, but rather to enable equitable, open, international access to the genomic data for all organisms.

Nonetheless, each member database has implemented some level of automated screening to ensure that minimum quality requirements are met for each new submission (9). In practice, data are accepted into INSDC databases without independent authentication or “validation,” and there is no requirement for reference specimens or source materials to be preserved or identified. Notably, the primary metadata standard used by INSDC member databases—the Minimum Information About Any (X) Sample (MIxS)—treats references for biomaterials (ref_biomaterial) and source material identifiers (source_mat_id) as optional or recommended fields (10–12). Thus, provenance (e.g., the documented origin and history of the data) and authenticity (e.g., confidence that data are genuine and untampered) are largely taken on trust. Unfortunately, while the genomics community generally operates

under the assumption that most researchers submit correct, well-documented sequences with clear sample (i.e., BioSample) metadata, this assumption is proving increasingly problematic (4–6, 13–16).

The absence of enforced data provenance and authenticity standards for both biological materials and experimental methods has led to the widespread practice of researchers contributing only the minimum information needed for each submission. Researchers worldwide deposit genomic data and associated metadata with varying degrees of quality control, often using inconsistent nomenclature and metadata formatting conventions. As expected, human-in-the-loop curation of these databases has not kept pace with their explosive growth. As a result, deposited data are rarely updated or corrected after submission, and the aggregate quality of sample and experimental metadata has degraded over time. Third-party curation and annotation systems, which have been proposed as solutions to this issue (17), have urged NCBI to allow a community-curated annotation process where domain experts could add or correct annotations on sequences. Unfortunately, this proposal was not embraced by INSDC member databases, and NCBI's own limited implementation—the Third Party Annotation [TPA] system—was retired in 2025, reportedly due to excessive technical complexity (18). As a result, public genomics repositories contain large numbers of entries that are incomplete (19–22), mislabeled (23, 24), of poor quality (25, 26), or outright fraudulent (27, 28). This significant but underappreciated problem is likely the result of a combination of insufficient minimum requirements for depositing data, poor scientific rigor, and in some cases intentional misconduct or obfuscation. Collectively, these deficiencies impact scientific reproducibility (7, 29–32), public health (5, 33, 34), and biosecurity (35–37).

Below, this manuscript reviews the specific issues surrounding the reuse, authentication, and provenance of genomic data, especially as they relate to descriptors for source materials, sample handling, chain of custody, sample processing, and bioinformatics methods. This work closes by arguing that the genomics community does not need to reinvent provenance standards, but must enforce and operationalize existing standards more consistently, preserve physical reference materials where possible, and prioritize targeted retrospective curation of high-value legacy records, especially those used as references, in clinical or regulatory contexts, or in widely reused AI/ML training and benchmark datasets.

Missing or Mislabeled Sample Metadata

A significant challenge in public genomic repositories is the pervasive lack of essential sample metadata, which are critical for interpreting and integrating genomic sequences

across nearly all potential downstream application areas. For pathogen genomics data, these gaps create real-world risks for public health and biosecurity. For example, a 2021 U.S. FDA-led study highlighted how key fields like collection date, geographic origin, host, and other contextual details are widely absent or incomplete in GenBank’s BioSample and GenomeTrakr pathogen surveillance databases, largely because MIxS guidelines make inclusion of this information optional at the point of submission (5). Their analysis of *Salmonella*, *Escherichia coli*, and *Listeria* genomes found that of the 550,270 genome assemblies inspected, ~25% of the entries were missing at least one crucial piece of metadata needed for public health surveillance and epidemiology. Furthermore, there was a significant disparity between samples identified as “clinical” vs. those from food or environmental sources: clinical samples were missing these metadata in 134,133 (~49%) of the records, whereas only 1,823 samples (0.7%) originating from food and environmental surveillance efforts had missing data. This suggests a meaningful disparity in awareness in the importance of these data among professionals operating in clinical laboratories and highlights an opportunity to improve training for scientists responsible for data collection and submission.

Our own analyses of publicly available bacterial genome assemblies also revealed significant metadata deficiencies (38). We examined 2,701 bacterial RefSeq genome assemblies labeled as “ATCC” reference strains and found that nearly half were missing one or more crucial metadata fields (or contained non-informative values) describing how the data were produced and processed. Furthermore, ~40% of records failed to report the sequencing technology or bioinformatics methods used to generate the assembly, and over 99% lacked any “isolate” description or source information associated with the originating BioSample. Even the “relation to type material” field (which indicates if a sequence comes from an official bacterial type strain) was empty in 38% of cases, making it difficult to discern whether a given genome sequence corresponds to the canonical type strain reference material for a species or instead represents a derivative, unverified laboratory strain labeled with the same strain name. In addition, among the assemblies examined that were generated from ATCC reference strains, internal ATCC records indicated that only ~15% of sequencing laboratories had obtained the organism directly from ATCC’s culture collection. The remaining ~85% presumably sequenced lab-to-lab clones or derivatives obtained through informal strain sharing between laboratories, a practice that effectively erodes one of the core purposes of international culture collections—namely, serving as a central controlled source of materials (39–41).

Similarly for eukaryotic data, a 2021 study by Buckner et al. (4) found that the majority of eukaryotic genetic sequences in GenBank are not linked to any preserved voucher specimen or culture collection record, despite prior calls to address this gap (42). The authors

identified a lack of physical traceability, errors in source material information, and the absence of clear descriptions on how the data were produced as collectively posing a “serious impediment” for researchers seeking to confirm that a DNA sequence truly corresponds to the species or strain. They provide cautionary examples in which mislabeled or contaminated sequences entered public databases and remain uncorrected, misleading research until errors are uncovered either unexpectedly or through labor-intensive forensic analyses (43). More recently, Crandall et al. (21) also found that most (87%) genome-scale genetic diversity data in GenBank lack the spatiotemporal metadata needed for reuse. Crandall’s “datathon” further demonstrated a persistent issue: even when metadata can be recovered, it often requires substantial effort involving manual curation due to gaps in data formatting and structure. Moreover, metadata recoverability declines rapidly with age (~13.5% loss per year), reinforcing the need to capture and curate metadata at the point of submission rather than attempting to “fix it later.” These critiques are also consistent with findings from Toczydlowski et al. who found, among more than 300,000 BioSamples tied to Sequence Read Archive (SRA) records relevant to biodiversity research, only ~13% include time and location information, substantially hampering the reuse of these data at scale (20).

Unfortunately, missing or incomplete metadata issues are not confined to genomic reference sequences but also significantly affect other important genomics databases such as NCBI’s Gene Expression Omnibus (GEO) data repository. In 2025, Huang et al. (22) systematically assessed how missing and inconsistently shared metadata in GEO undermine data reuse and reproducibility: across 253 studies (>164,000 samples) they found that ~25% of critical metadata were omitted, significantly hampering comparative analyses across studies. Furthermore, only 11.5% of these studies fully shared the assessed phenotypes, meaning most records lack the basic descriptors needed to interpret datasets or match samples across studies. The authors explicitly note that metadata fields for sample source (and other similar contextual fields) were often missing or inconsistently reported in both SRA and GEO, thereby limiting reusability even when raw data are deposited.

Another study reached a similar conclusion for microbiome-oriented INSDC deposits: despite broad awareness and uptake of MIxS, metadata associated with most samples remained poorly standardized, nonstandard fields introduced substantial variation, and harmonization across studies was often difficult or impossible without manual curation (44). This issue further exacerbates challenges in downstream analytics, including machine learning and AI applications, revealing fundamental shortcomings in how bioinformatics provenance is captured, structured, and maintained.

Bioinformatics Provenance Gaps

Although NCBI repositories such as SRA, BioSample, and GenBank implement structured submission schemas—and in some domains incorporate community standards such as MlxS—these frameworks still capture only a subset of the methodological record. In SRA, technical metadata are organized around the STUDY, SAMPLE, EXPERIMENT, and RUN objects; the EXPERIMENT and RUN records capture items such as library preparation, sequencing strategy, layout, and instrument model (45). However, NCBI also states that most descriptive information is captured at the EXPERIMENT level and depends on submitter-provided Title and Description text. BioSample likewise maintains recognized attributes and package structures but explicitly permits submitters to provide any number of custom attributes. GenBank genome submissions require a limited set of assembly metadata, including assembly method, program version or date, approximate coverage, and sequencing technology, while other informative details such as polishing method or reference-guided status remain optional. Taken together, these repository schemas provide a scaffold for deposition, but they do not constitute a single, consistently enforced, machine-readable standard for end-to-end sequencing and bioinformatics provenance.

This limitation is not merely theoretical. A 2015 study examined SRA protocol annotation and found that only ~4% of studies contained annotations for queried preparatory protocol steps, while ~74% of studies had omitted details for those steps; the authors concluded that the current level of annotation inhibits systematic studies of protocol bias and significantly weakens comparative analysis (46). Given the sustained rapid growth of SRA over the last decade, it remains unclear whether any effort has been made to retrospectively improve these metrics for historical samples. At the sample-metadata layer, MetaSRA was developed precisely because SRA-linked BioSample metadata were difficult to use at scale: both property names and values were described as non-standardized and created at the discretion of submitters, resulting in synonyms, misspellings, abbreviations, and free-text phrases that resist large-scale computational reuse (47). Klie and colleagues reported that most SRA samples were missing metadata across several categories and that user-defined fields dominated the metadata landscape. In their dataset, only ~22% of values were paired with BioSample-defined attributes. Another 2019 study found the same pattern in BioSample more broadly: 85% of records used the Generic package, which imposes very few required attributes, and 15% of attribute name-value pairs used *ad hoc* names outside the BioSample dictionary (48). The authors further concluded that most field names and values were not standardized or controlled; many distinct labels were used to represent the same underlying concept, and these inconsistencies impede search and reuse.

A central issue across public databases is that metadata are frequently provided only at the experiment level rather than the sample level. The studies cited above emphasize that missing sample-level metadata restricts credibility and can make secondary analysis a substantial challenge (i.e., replication and robust reanalysis require per-sample descriptors). These findings indicate that public archives often preserve raw sequence files more reliably than they preserve the contextual and methodological metadata needed for robust cross-study interpretation, replication, and reuse.

The problem becomes even sharper when the focus shifts from sample description to computational provenance. Metadata standards may record platform, library, and assembly-summary fields, but they do not generally require a standardized disclosure of the complete analytical environment: software versions, parameter settings, reference databases, filtering thresholds, contamination-screening logic, branching decisions, or validation procedures (49). This demonstrates why complementary provenance frameworks such as BioCompute are needed as a standard for communicating high-throughput sequencing workflows, validation, and reproducibility (50). Workflow-provenance studies confirm that incomplete documentation of workflow requirements can lead directly to failed re-execution in new environments (51). Therefore, the methodological opacity of public genomics archives is not simply a metadata nuisance; it is a structural trust problem. When sequencing and bioinformatics methods are inconsistently recorded, relegated to free text, or omitted entirely, independent verification becomes harder (and often impossible), comparative reuse becomes more fragile, and the evidentiary value of archived genomic data is weakened. These gaps directly shape whether genomic analyses can be reproduced, meaningfully reanalyzed, or trusted as evidence

Scientific Misconduct and Data Fabrication

Scientific fraud, including data fabrication, falsification, and plagiarism, should be of significant concern for anyone working in the genomics, bioinformatics, genetics, AlxBio, or digital biology space—especially those working with data sourced from public data repositories. Unfortunately, however, a high degree of presumed trust is placed in data sourced from these databases, despite ample signs that these data should be treated with caution. While most scientists are honest, surveys and meta-analyses indicate a non-trivial minority engage in unethical practices that collectively create significant risks for the entire global genomics research community. A large-scale systematic survey and meta-analysis of prior work on scientific fraud revealed ~2% of scientists in the U.S. and EU admitted to fabricating or falsifying data at least once within the three years prior to being surveyed (52). A subsequent study surveyed 6,813 scientists and found ~4% admitted to fabricating or falsifying existing data one or more times in their careers, largely driven by pressures to

publish (53). The problem may be even more widespread: in 2023, an estimated 5.8% of all biomedical research papers published globally were “true fakes” and reported completely fabricated results (54). The issue of fabricated publications has been repeatedly raised by Jennifer Byrne and colleagues, most recently in their analysis of 2.6 million cancer research publications spanning 1999 to 2024 where ~9.9% of all publications were flagged as suspicious and likely originating from “paper mill” publishers (55).

Raw genomic data are typically structured as plain-text files (i.e., .fasta, .fastq, .gbk, etc.), making them amenable to computational modeling. As a result, generative approaches for simulating realistic genomic sequence data in silico have been an active area of research in computational biology for decades (56–61). These tools are generally used for benchmarking and testing newly developed bioinformatics algorithms at scale, obviating the need for laboratory experiments. Importantly, however, these same tools can also be used to generate fabricated genomic data or to augment existing datasets with fabricated sequences to deliver a specific result needed for publication. Pairing such datasets with falsified metadata would be trivial. Taken together, these data could be submitted to public genomics databases to support a fabricated publication, which would result in the poisoning of public database integrity and accuracy. Unfortunately, there have been documented instances where falsified genomics data has been found in these databases.

The earliest known example of falsified sequencing data associated with peer-reviewed publications occurred in the late 1990s (62). The associated data were later “suppressed” from GenBank after the publications were retracted, but not before being cited 27 times in the literature. In another study, researchers found falsified mitochondrial DNA sequences to support a series of publications and a fabricated phylogenetic tree of sparrow hawks. This was discovered serendipitously after researchers had difficulty identifying the source of voucher specimens for the original (falsified) data and made open calls for improved curation and stronger enforcement of authenticity checks for mitochondrial DNA sequences in public databases (13, 63, 64). Interestingly, these data remain labeled as “UNVERIFIED” in GenBank to this day, leaving the potential for their successful retrieval using common tools like BLAST despite being entirely false. A third example of genomics data misconduct involved falsified and intentionally mislabeled gene expression microarray data put forth as evidence of genomic signatures suitable for companion diagnostics in chemotherapeutics (65). At least two groups independently failed to replicate these results, but the data submitted to GEO were not removed until a third group published a separate peer-reviewed “forensic bioinformatics” publication calling for the retraction of the original paper and data (66). In a fourth case, a scientist falsely claimed successful integration of a therapeutic gene in mice, fabricating DNA sequencing data for CRISPR integration sites and related PCR assays. A 2010 NIH investigation resulted in the retraction of three papers and the

cancellation of four grants (67). In a fifth case, DNA methylation sequencing data was falsified by an early-career scientist investigating endocrine disruptors in animal models. After the results could not be replicated, another NIH Office of Research Integrity (ORI) investigation revealed the original results were fabricated, resulting in the retraction of the corresponding paper and removal of the data from public databases (68). Importantly, all the above instances involved discovery of the fabrication only after independent groups attempted to replicate the results, which then triggered investigations once those attempts failed. In all these cases, the removal or suppression of the data did not occur until years after the data were originally deposited and, in some cases, after the original fabricated work was cited by hundreds of separate publications.

The examples above represent a small number of readily confirmed cases in which fabricated or falsified genomics or omics data were deposited into public repositories. While the number of documented cases may seem limited, it should be interpreted cautiously given how little systematic work has been directed at detecting fabricated genomic data within public archives. It's also important to consider these findings within the broader context of scientific fraud in general. By the most conservative estimate above, the impact of even 2% of genomics scientists falsifying data would have an enormous detrimental impact on our public data repositories.

Very limited research has been directed specifically at detecting fraudulent genomic data that may already exist in INSDC databases (as opposed to published works). Only a single published study has intentionally attempted to detect fraudulent sequencing data directly (69). Ironically, the authors note they were not successful in identifying fraudulent data in GenBank associated with retracted papers, not due to the absence of the data, but instead largely due to gaps in sample and depositor attribution metadata for these data, which made it extremely difficult to properly assess the veracity of data that may be tied to public databases. Separately, Bouadjenek et al. developed a means to assess the quality of sequencing records as a whole (i.e., both the sequence and associated metadata) using a combination of machine learning models and iterative information retrieval strategies (70). They report ~25% of the sequencing records investigated were “suspicious”, although they did not specifically identify these as potentially fabricated and instead frame these as data quality oversights by the submitters. Thus, despite the relative ease with which fabricated genomic data could be produced at scale, the degree to which intentionally falsified data is currently poisoning public databases remains largely unanswered.

Reproducible Does Not Mean Correct

Irreproducibility is a well-recognized crisis in science, and deficiencies in genomic data quality and metadata completeness materially undermine reproducibility and reuse of genomic data (71–73). The impacts of genomic data contamination (9, 25, 33), incomplete metadata (22, 30, 74), sample-quality imbalance (75), and weak data provenance (38, 51) have been well documented. If published genomic datasets are contaminated, mislabeled, or otherwise unreliable, other scientists may spend extended periods of time chasing false leads or trying and failing to replicate results found in existing published datasets. Worse, they may computationally replicate results from otherwise compromised data and unknowingly reach the same (incorrect) conclusions, a scenario that is often overlooked by early-career scientists who may assume that public datasets are correct (76). Accordingly, the central question is not only whether a genomic result can be reproduced, but whether the data and methods underlying that result are sufficiently authentic, traceable, and well-described to justify confidence in its correctness. Under this model, reproducing prior results then tests both the technical aspects of the replication experiment and the biological validity of the results.

Conceptually, reproducibility, replicability, and correctness should not be treated as interchangeable. Reproducibility demonstrates that a result can be regenerated using the same data and analysis; replicability asks whether similar conclusions are obtained when a study is repeated independently; correctness concerns whether the underlying biological claim is true (72, 76). These distinctions matter in genomics because technically reproducible workflows may still yield false biological inferences if the underlying data are contaminated, mislabeled, incompletely described, or interpreted against flawed reference resources. In this sense, reproducibility is necessary for trust, but it is not sufficient. Importantly, archival deposition in repositories such as BioSample, BioProject, GenBank, or SRA should not be mistaken for independent validation. As extensively discussed above, these submitter-driven archives are designed primarily to preserve and disseminate data, not to certify the correctness of the biological claims associated with those data.

An additional concern is error propagation through shared databases and analytical infrastructure. Once a contaminated, misclassified, or otherwise defective record enters a public repository, it can be reused in downstream annotation pipelines, comparative genomics studies, taxonomic classifiers, benchmark datasets, meta-analyses, and myriad other reuse cases (73). Subsequent investigators may then obtain concordant results not because the underlying claim is true, but because they are relying on the same compromised inputs. This creates a form of circular confirmation in which “reproducible wrongness” is reinforced by repeated reuse. The problem is amplified by the fact that reference databases

themselves are not error-free: large-scale contamination, taxonomic misassignment, and sequence-content errors have all been documented in public sequence resources (23, 26, 34, 77). Thus, even technically rigorous analyses may produce stable but incorrect answers when the shared reference layer is flawed.

Finally, it is important to recognize that “correctness” in genomics is not a static notion, but a moving target that shifts as scientific understanding advances. However, individual genomics datasets are largely static records that are seldom retrospectively updated or systematically re-analyzed after the data are submitted—even when new evidence reveals prior classifications, annotations, or assemblies are no longer accurate, e.g., the recent work of Nguyen and others (78–80). Consequently, older genome entries produced under outdated taxonomic frameworks, legacy methods, or initial assumptions often persist uncorrected in these databases. When subsequent studies unknowingly reuse such legacy data, they may propagate obsolete interpretations and effectively contaminate modern analyses, reinforcing a cycle of wrongness in which results appear consistent but rest on outdated information. In essence, reproducibility alone cannot guarantee true biological validity if reference databases fail to keep pace with new scientific insights: one can faithfully replicate past findings yet still fall out of step with current scientific truth. This reality underscores that data stewardship involves more than simply collecting and providing data. An ideal genomic data ecosystem would include continuous curation and versioning of reference records, near real time dissemination and verification of updated annotations or reclassifications, and richer, standardized metadata to enable downstream researchers to reinterpret legacy datasets as knowledge evolves. Although implementing such practices is challenging under current community-submission models (especially given the frequent lack of comprehensive metadata), strengthening these stewardship mechanisms is increasingly crucial to ensure that shared genomic resources remain as up to date, accurate, and trustworthy as the science they underpin.

Metadata quality and computational provenance determine whether such wrongness can be recognized and investigated after the fact. Missing sample-level metadata, non-standardized descriptors, and incomplete method reporting do not merely reduce convenience; they directly limit the ability of other investigators to distinguish true biological variation from contamination, sample-quality imbalance, sample mislabeling, or analytical artifacts (5, 6, 22, 44, 48, 74). Rajesh et al. (2021) found that substantial metadata are lost between publications and public repositories, while Huang et al. (2025) showed that critical metadata omissions remain widespread across public omics studies. At the computational level, incomplete documentation of workflow requirements, software versions, parameters, and execution environments can prevent meaningful reanalysis or even successful re-execution. This is precisely why provenance-focused efforts such as BioCompute emerged

(50): not because reproducibility was solved, but because the archives and publications alone often fail to capture enough information to interpret what was actually done. Lastly, these issues have real economic consequences as well, by some estimates as high as \$28 billion in research funding is wasted annually in the U.S. alone (81).

Challenges for Digital Biology

Ultimately, the issues described above extend into the integrity of AI-enabled biology, going far beyond classical reproducibility and data provenance issues. Public genomic data are increasingly reused for model training, benchmarking, and reference-set construction. If those data are missing material origin information or have unresolved contamination, metadata inconsistencies, provenance gaps, or duplicated and mislabeled records, the resulting models may appear performant while encoding hidden biases or failure modes inherited from the training corpus (7). Recent standards-focused reviews have explicitly noted that data-quality variability, inconsistent metadata, and reuse barriers in genomic archives could affect emerging AI models under development (7, 22, 44).

For digital biology, data provenance and structured metadata are a matter of enabling infrastructure. Recent advances in AI-enabled biology already illustrate this, with AlphaFold representing the clearest positive example. In the original AlphaFold paper, it was stated that the model was possible because it could train to high accuracy using supervised learning on Protein Data Bank (PDB) data (82). Subsequent PDB reviews have been even more explicit: rigorously validated and expertly biocurated PDB structures are considered a gold-standard resource that has made AI/ML prediction of protein structures possible (83). This did not happen because the PDB was merely a large, comprehensive database, but rather because it has maintained a very high level of rigorous deposition standards, formal validation, expert biocuration, and public validation reports since its inception (84). In short, AlphaFold was enabled not only by architecture and compute, but by decades of disciplined, human-in-the-loop data stewardship.

In contrast, genome-scale foundation models are being developed under much less favorable data conditions. Evo 2 was trained on 9 trillion DNA base pairs from a “highly curated genomic atlas” (OpenGenome2), whose corpus was compiled from curated, non-redundant nucleotide data spanning bacteria, archaea, eukaryotes, and bacteriophages (85). Interestingly, OpenGenome2 was not based on NCBI’s RefSeq, but instead was based on GTDB (86), IMG/VR (87), and IMG/PR (88), with additional layers of automated curation applied in all cases to further quality-filter the resulting data used for building OpenGenome2 and training Evo 2. Likewise, the recently published Nucleotide Transformer model also did not simply ingest arbitrary public genomes; its pretraining was selected from

RefSeq on the basis of assembly quality and species diversity, yet significant manual curation of records was still needed for model development (89). Similarly, CellFM development was directly hampered by technical noise and batch effects across single-cell sequencing datasets, which required extensive standardized data cleansing of public datasets into unified formats before training (90). Another single-cell AI model being developed, SCimilarity, had the same challenges: pan-body analyses were hampered by dataset curation and harmonization challenges stemming from gaps in metadata and a lack of standardized pre-processing methodologies (91). A recent review of large single-cell genomics atlases reinforces these points (92). Thus, even in the emerging field of genomics foundation models, developers are already compensating for uneven public data quality and poorly documented metadata by imposing additional filters, human curation decisions, and label controls before training or evaluation.

Benchmarking and evaluation introduce an additional layer of risk. The Nucleotide Transformer authors explicitly note that Enformer's performance in their benchmark might be inflated because the model had originally been trained using different data splits, creating potential data leakage (89). A later benchmark of DNA foundation models reached a similar conclusion from another angle, showing that pretraining data composition and metadata accuracy materially affect downstream performance (93). Consistent with these concerns, Joeres et al. recently introduced DataSAIL, a framework aimed at helping prevent overly optimistic benchmark results driven by information leakage rather than true model generalization (94).

Establishing Reference Source Materials

In 2018, ATCC launched the Enhanced Authentication Initiative (EAI) to perform whole-genome sequencing on 250 microbial strains that serve as critical reference materials for clinical microbiology and were known to have significantly divergent genome references in public databases. At a high level, the goal of the EAI was to establish ground truth for these microbes' genomes by sequencing materials directly from ATCC's reference collection. This program was later expanded in 2019 with the launch of the ATCC Genome Portal (AGP, <https://genomes.atcc.org>), which has the implicit goal of creating high-quality genomic reference data for all materials held within ATCC's living collection (95). Today, the AGP is composed entirely of genomic data produced in-house by ATCC for 7,000 microbes and nearly 1,000 human and animal cell lines (as of June, 2026). All of the data were generated directly from source materials in ATCC's collection, using both Oxford Nanopore Technologies and Illumina sequencing technologies, and subsequently layered with expertly curated provenance and sample source data, genome annotations, resistome data,

methylation data, transcriptomics data, and genomic variant data for thousands of ATCC microbes, human and animal cell lines, and viruses.

ATCC's initiative was not the first to establish clear provenance between materials and the genomic data representing them. Early examples of this can be found in the Cancer Cell Line Encyclopedia (CCLE), which was initiated in 2006 as a collaboration between Novartis and the Broad Institute and released in 2012 (96). CCLE included multiomics data for ~1,000 cell lines commonly used across the biopharma and cancer research domains and aimed to create a digital atlas for benchmarking and exploratory studies. The provenance of the materials used for CCLE, however, has come into question multiple times since the initial landmark publication. Significant genetic and pharmacogenomic discrepancies between its database and other major repositories, such as the Sanger Institute's, have been documented by multiple independent studies (97–99). An early study, which was later openly debated, identified issues with passage history and secondary distribution as potential sources of differences between CCLE and the Cancer Genome Project (CGP), where a lack of direct traceability to certified lots (like those from ATCC) may have contributed to mislabeled lines or irreproducible drug-response data. A later study identified 106 "identical" cell lines shared between these institutes that showed only a 57% overlap in mutations, suggesting that many lines have evolved into genetically distinct strains rather than being identical reagents and, more importantly, that these materials had not been directly sourced from a common biorepository or culture collection (100). The general issues around cell line authenticity and the importance of provenance specifically were later reviewed in 2018 by Hynds et al. (101). A systematic comparison of ATCC's own 'omics data for the ~1,000 cell lines currently included in the AGP against those from CCLE and CGP is underway.

In microbiology, similar efforts have been undertaken. For example, a consortium was led by the U.S. FDA to create the FDA-ARGOS BioProject where human pathogen "reference strains" were sequenced using "regulatory-grade" quality criteria (102). This cohort of data was later expanded to include genome assemblies for 1,428 microbes. The FDA-ARGOS database introduced additional hurdles to its utility, however, by intentionally renaming known strains, some available in culture collections, with new "FDAARGOS_#####" strain designations. It is also notable that at least 83 (~5.8%) of the assemblies in the FDA-ARGOS BioProject (PRJNA231221) have subsequently been marked as "atypical" by NCBI's automated systems either for potential source material ambiguity, taxonomic misclassification, or contamination (e.g., ~74 have greater than 5% contamination by CheckM using family-specific markers) and as a result, are suppressed genomes. Lastly, since many of the isolates used for FDA-ARGOS were sourced from non-public clinical microbiology laboratory collections, and the associated strain identifiers for publicly

available strains were renamed, replicating the results of FDA-ARGOS even for a small number of isolates remains a substantial challenge. Collectively, this limits the ability of the broader research community to associate these reference genomes with existing type-strains or clinical reference strain identifiers and therefore limits the utility of the data.

In a separate effort, the National Collection of Type Cultures (NCTC) created the NCTC3000 collection where 2,234 whole-genome assemblies were produced using only PacBio data (103). The effort primarily focused on bacterial pathogens, with an emphasis on type strains and historical isolates from the Public Health England's NCTC. In a similar vein, the Japanese Collection of Microorganisms (JCM) carried out whole-genome sequencing of 351 authenticated prokaryote strains for which no prior reference genome was available, an effort that helps expand the overall diversity of microbial genomic data for otherwise neglected organisms (104). Like ATCC's AGP initiative, the NCTC3000 collection is derived from authenticated strains and culture collection numbers (NCTC IDs) link directly to stored vials in the biobank for each assembly.

A similar effort was undertaken using a democratized consortium approach by the World Federation of Culture Collections' (WFCC) 10K genomes project (105), which currently is ~20% complete (e.g., see BioProject #PRJDB9057). Importantly, while the WFCC project outlines protocols and methods for participating sequencing centers, the metadata captured, sourcing of materials, sample processing, and handling of the isolates was not standardized and left to submitting researchers to decide. This may be one of the underlying reasons why ~20% of the assemblies have >5% contamination scores (e.g., see CheckM values using family level marker indices as reported in PRJDB9057). Furthermore, some of the physical source materials used for each assembly were deposited into culture collections by researchers after the data were made available, which creates gaps in the tracking of sample handling. In fact, 109 of the assemblies in this BioProject are "atypical" and not attributable to a known collection. Thus, while this initiative aims to create a linkage between genomic data and physical strains, in practice the provenance of the data has created a high degree of uncertainty in the accuracy of the results. This is in contrast to the "sourced from the collection first" approach taken by ATCC, the JCM, and the NCTC.

As the AGP has grown, and as our own methods have improved as new technologies and bioinformatics tools have emerged, so too has the quality and comprehensiveness of our data. Furthermore, these changes have demanded re-evaluation of how datasets are linked to one another, what metadata are captured, what criteria are necessary to establish provenance and trust, and how these can be provided to end-users of the AGP. This is an iterative exercise in the case of the AGP as community needs change and ATCC's capabilities increase. This "living database" model is one of the key differentiators between ATCC's

initiative and (at scale) it is likely not feasible for massive community-driven data archives like RefSeq, but it is plausible for culture collections to achieve.

Taken together, these efforts illustrate that trustworthy genomic reference data are enabled not by sequencing scale or technological sophistication alone, but by the sequencing of materials whose physical provenance, handling history, and custodial chain are well defined at the outset. Initiatives such as the AGP and NCTC3000 exemplify a “collection-first” model, in which genomic data are generated directly from authenticated, traceable source materials. In contrast, sequence-first or community-aggregated efforts often struggle to retrospectively establish provenance, even when data quality metrics appear strong. This distinction has important implications for how genomic reference resources are constructed, curated, benchmarked, and trusted, particularly as such datasets increasingly underpin clinical, regulatory, and AI-driven biological analyses.

Discussion

Standards for genomics metadata and provenance are not the limiting factor: implementation and enforcement of those standards is. Community frameworks such as MxS/MIGS (10, 11, 106), Darwin Core (e.g., GGBN) (107, 108), public health specifications (e.g., PHA4GE) (109, 110), and workflow provenance standards (e.g., BioCompute) (50) all make the same underlying claim: genomic sequences are only meaningfully reusable when they remain connected to (i) a traceable physical sample or specimen, (ii) the key contextual descriptors of how that sample was collected and processed, and (iii) the computational methods used to transform raw reads into the deposited record. Yet in major public archives, critical provenance fields remain optional, inconsistently populated, and weakly validated, so records frequently preserve the sequence more reliably than the evidence needed to interpret, replicate, reproduce, or validate it (22, 48, 74).

A core gap is that the provenance of sample origins, handling, and primary sample processing methodologies are not mandated as a first-class, “metadata as infrastructure” requirement. Instead, it is applied inconsistently through a patchwork of different labeling schemas and requirements. Even when standards define “source material identifiers” or voucher linkages, public repositories often accept submissions without persistent connections to the originating biomaterial, leaving no reliable chain-of-custody and no practical route to obtain the original material for confirmation (5, 6, 42, 43, 48, 71, 74). This is not a niche concern: across multiple domains, metadata incompleteness and inconsistency remain widespread, and the burden of reconstruction falls to downstream end users rather than depositors.

A second gap is that the provenance of the computational methods used is not captured in a consistently machine-readable way (46, 47, 111). Archive schemas may store platform and basic run metadata, but critical details needed for replication (i.e., specific software versions, parameters, reference resources, filtering/contamination screening decisions, and validation logic) are often relegated to free text or (more commonly) omitted entirely. The existence of separate provenance standards like BioCompute reflects this reality: they arose because repositories and papers alone frequently fail to preserve enough workflow detail for reliable re-execution or regulatory-grade traceability (50, 112). Notably, BioCompute Objects have been identified formally by the FDA as an acceptable model for tracking the provenance of submissions for regulatory decisions (113). Similarly, the FAANG and ENCODE metadata models are examples of data standards requiring rich metadata for sample and data provenance. FAANG explicitly framed rich sample and experimental metadata as a core goal (114), and ENCODE organizes assay and computational metadata as structured schema objects (115). Yet, despite these major efforts to develop accessible, open interoperability standards that will work broadly with genomic data, these models have not been integrated into major public data repositories and remain largely parallel standards to the arguably less rigorous MIxS standards widely in use.

Finally, the public archive ingestion model prioritizes scale over verification. For example, GenBank is built primarily from direct submissions by researchers and sequencing centers, underscoring that deposition is not intrinsically gated by peer review or independent validation. Data depositors can update records after they are released with links to related peer-reviewed publications, but in practice this is not a guarantee. It's interesting to note the historical contrast of this approach to Goad's original intent for the Los Alamos DNA Data Bank, wherein all records would be labeled as "provisional" unless they were independently validated (1). Thus, public sequence repositories function as enormous data lakes with uneven provenance rather than curated reference resources: they maximize openness and volume, but provide limited guarantees about sample origin, authenticity, handling history, or computational traceability. This structural gap creates an ongoing challenge for method validation studies, complicates comparative genomics, imposes obstacles for genomics-driven public health responsiveness, and hampers the development and training of high-accuracy AIxBio models for digital biology applications.

When Walter Goad proposed the first centralized DNA sequence repository in 1979, he stressed the importance of collecting "complete information on DNA origin," including the source organism and even original lab records for verification. Goad's vision included archiving gel electrophoresis autoradiograms and detailed sample information on microfilm alongside sequence data to ensure that future users could evaluate a sequence's authenticity and quality (1). In many ways, current calls for data provenance echo those

early recommendations. The fact that the community is still grappling with missing metadata suggests that provenance and authenticity have been chronically undervalued for decades.

Public genomics repositories have been indispensable for accelerating discovery, but their current “archive-first” operating model leaves a growing trust gap: sequences are often preserved more reliably than the evidence needed to interpret, reproduce, or validate them. As datasets scale and downstream reuse expands, especially for AIxBio applications, missing or inconsistent provenance (sample origins, handling history, and chain-of-custody) and incomplete computational provenance (software, parameters, reference resources, and QC decisions) increasingly translate into irreproducible findings, silent error propagation, and exploitable opportunities for fabrication or misuse. As a result, the life sciences community is now investing significant effort to retrofit these databases with better metadata at considerable costs, an endeavor that will be crucial for the next era of data-driven biology.

In addition, while prevention is essential, the scale of historical metadata gaps in public repositories cannot be ignored either. Strategic, targeted retrospective curation prioritized around high-value reference records (e.g. clinical and regulatory use cases) and widely reused benchmarking datasets will be necessary. Doing so will substantially aid in mitigating existing risks, even if comprehensive remediation is infeasible. Efforts to curate legacy data should therefore be treated as an infrastructure investment and aligned with clear incentives, automated where possible, and guided by realistic expectations about what can and cannot be recovered after the fact.

In conclusion, the genomics and digital biology community does not need to reinvent standards; the existing standards the genomics community has already created must be enforced in ways that reduce friction for submitters while materially improving interoperability, traceability, and reuse. Critically, physical reference samples and the metadata that describe their origin, handling, and transformation should be treated as shared scientific infrastructure that is proactively preserved, persistently identified, and curated with the same long-term intent as reference genomes or computational resources. Culture collections and biorepositories can uniquely bridge the physical-to-digital divide by anchoring verifiable “digital twins” to authenticated source materials (41, 43, 95), while data repositories and journals can drive adoption through raising the minimum deposition requirements for metadata, better submission tooling, and visible incentives for strong provenance. The genomics community has long championed open data; the next step is to champion *trustworthy* data sharing so that the future of digital biology and AI-enabled biology is built on traceable, defensible, and reusable genomic records, rather than on an assumption of trust.

References

1. W. B. Goad, "Proposal to establish a national center for collection, and computer storage and analysis of nucleic acid sequences." (Proposal, University of California. Los Alamos Scientific Laboratory, 1979); <https://hdl.handle.net/10822/556965>.
2. E. Jordan, C. Carrico, DNA Database. *Science* **218**, 108–108 (1982).
3. National Center for Biotechnology Information, GenBank and WGS Statistics, *GenBank and WGS Statistics* (2026).
<https://www.ncbi.nlm.nih.gov/genbank/statistics/>).
4. J. C. Buckner, R. C. Sanders, B. C. Faircloth, P. Chakrabarty, The critical importance of vouchers in genomics. *eLife* **10**, e68264 (2021).
5. J. B. Pettengill, J. Beal, M. Balkey, M. Allard, H. Rand, R. Timme, Interpretative Labor and the Bane of Nonstandardized Metadata in Public Health Surveillance and Food Safety. *Clinical Infectious Diseases* **73**, 1537–1539 (2021).
6. A. Rajesh, Y. Chang, M. S. Abedalthagafi, A. Wong-Beringer, M. I. Love, S. Mangul, Improving the completeness of public metadata accompanying omics studies. *Genome Biol* **22**, 106, s13059-021-02332-z (2021).
7. I. Keenum, S. A. Jackson, E. Eloë-Fadrosh, L. M. Schriml, A standards perspective on genomic data reusability and reproducibility. *Front. Bioinform.* **5**, 1572937 (2025).
8. G. Cochrane, I. Karsch-Mizrachi, T. Takagi, I. N. Sequence Database Collaboration, The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res* **44**, D48–D50 (2016).
9. A. Astashyn, E. S. Tvedte, D. Sweeney, V. Sapojnikov, N. Bouk, V. Joukov, E. Mozes, P. K. Strobe, P. M. Sylla, L. Wagner, S. L. Bidwell, L. C. Brown, K. Clark, E. W. Davis, B. Smith-White, W. Hlavina, K. D. Pruitt, V. A. Schneider, T. D. Murphy, Rapid and sensitive detection of genome contamination at scale with FCS-GX. *Genome Biol* **25**, 60 (2024).
10. D. Field, G. Garrity, T. Gray, N. Morrison, J. Selengut, P. Sterk, T. Tatusova, N. Thomson, M. J. Allen, S. V. Angiuoli, M. Ashburner, N. Axelrod, S. Baldauf, S. Ballard, J. Boore, G. Cochrane, J. Cole, P. Dawyndt, P. De Vos, C. dePamphilis, R. Edwards, N. Faruque, R. Feldman, J. Gilbert, P. Gilna, F. O. Glöckner, P. Goldstein, R. Guralnick, D. Haft, D. Hancock, H. Hermjakob, C. Hertz-Fowler, P. Hugenholtz, I. Joint, L. Kagan, M. Kane, J. Kennedy, G. Kowalchuk, R. Kottmann, E. Kolker, S. Kravitz, N. Kyrpides, J. Leebens-Mack, S. E. Lewis, K. Li, A. L. Lister, P. Lord, N. Maltsev, V. Markowitz, J. Martiny, B. Methe, I. Mizrahi, R. Moxon, K. Nelson, J. Parkhill, L. Proctor, O. White, S. A. Sansone, A. Spiers, R. Stevens, P. Swift, C. Taylor, Y. Tateno, A. Tett, S. Turner, D.

- 676 Ussery, B. Vaughan, N. Ward, T. Whetzel, I. San Gil, G. Wilson, A. Wipat, The
677 minimum information about a genome sequence (MIGS) specification. *Nature*
678 *Biotechnology* **26**, 541–547 (2008).
- 679 11. P. Yilmaz, R. Kottmann, D. Field, R. Knight, J. R. Cole, L. Amaral-Zettler, J. A. Gilbert, I.
680 Karsch-Mizrachi, A. Johnston, G. Cochrane, R. Vaughan, C. Hunter, J. Park, N.
681 Morrison, P. Rocca-Serra, P. Sterk, M. Arumugam, M. Bailey, L. Baumgartner, B. W.
682 Birren, M. J. Blaser, V. Bonazzi, T. Booth, P. Bork, F. D. Bushman, P. L. Buttigieg, P. S. G.
683 Chain, E. Charlson, E. K. Costello, H. Huot-Creasy, P. Dawyndt, T. DeSantis, N. Fierer,
684 J. A. Fuhrman, R. E. Gallery, D. Gevers, R. A. Gibbs, I. S. Gil, A. Gonzalez, J. I. Gordon,
685 R. Guralnick, W. Hankeln, S. Highlander, P. Hugenholtz, J. Jansson, A. L. Kau, S. T.
686 Kelley, J. Kennedy, D. Knights, O. Koren, J. Kuczynski, N. Kyrpides, R. Larsen, C. L.
687 Lauber, T. Legg, R. E. Ley, C. A. Lozupone, W. Ludwig, D. Lyons, E. Maguire, B. A.
688 Methé, F. Meyer, B. Muegge, S. Nakielny, K. E. Nelson, D. Nemergut, J. D. Neufeld, L.
689 K. Newbold, A. E. Oliver, N. R. Pace, G. Palanisamy, J. Peplies, J. Petrosino, L. Proctor,
690 E. Pruesse, C. Quast, J. Raes, S. Ratnasingham, J. Ravel, D. A. Relman, S. Assunta-
691 Sansone, P. D. Schloss, L. Schriml, R. Sinha, M. I. Smith, E. Sodergren, A. Spor, J.
692 Stombaugh, J. M. Tiedje, D. V. Ward, G. M. Weinstock, D. Wendel, O. White, A.
693 Whiteley, A. Wilke, J. R. Wortman, T. Yatsunenko, F. O. Glöckner, Minimum
694 information about a marker gene sequence (MIMARKS) and minimum information
695 about any (x) sequence (MlxS) specifications. *Nat Biotechnol* **29**, 415–420 (2011).
- 696 12. Genomic Standards Consortium, MIXS Checklists (2026).
697 <https://genomicsstandardsconsortium.github.io/mixs/#checklists>.
- 698 13. G. Sangster, J. A. Luksenburg, Scientific data laundering: Chimeric mitogenomes of a
699 sparrowhawk and a nightjar covered-up by forged phylogenies. *Biochemical*
700 *Systematics and Ecology* **96**, 104263 (2021).
- 701 14. D. A. Yarmosh, J. G. Lopera, N. P. Puthuveetil, P. F. Combs, A. L. Reese, C. Tabron, A.
702 E. Pierola, J. Duncan, S. R. Greenfield, R. Marlow, S. King, M. A. Riojas, J. Bagnoli, B.
703 Benton, J. L. Jacobs, Comparative Analysis and Data Provenance for 1,113 Bacterial
704 Genome Assemblies. *bioRxiv*, doi: 10.1101/2021.12.14.472616 (2021).
- 705 15. E. W. Sayers, M. Cavanaugh, K. Clark, K. D. Pruitt, S. T. Sherry, L. Yankie, I. Karsch-
706 Mizrachi, GenBank 2024 Update. *Nucleic Acids Res* **52**, D134–D137 (2024).
- 707 16. A. Carné, D. R. Vieites, M. P. Van Den Burg, In Vouchers We (Hope to) Trust: Unveiling
708 Hidden Errors in GENBANK 's Tetrapod Taxonomic Foundations. *Molecular Ecology* **34**,
709 e17812 (2025).
- 710 17. M. I. Bidartondo, Preserving Accuracy in GenBank. *Science* **319**, 1616–1616 (2008).
- 711 18. The International Nucleotide Sequence Database Collaboration, From January 2025
712 TPA-Exp and TPA-Inf submission types will no longer be accepted as new

713 submissions (03-09-2024) (2024). [https://www.insdc.org/news/from-january-2025-](https://www.insdc.org/news/from-january-2025-tpa-exp-and-tpa-inf-submission-types-will-no-longer-be-accepted-as-new-submissions-03-09-2024/)
714 tpa-exp-and-tpa-inf-submission-types-will-no-longer-be-accepted-as-new-
715 submissions-03-09-2024/.

716 19. O. K. Tørresen, B. Star, P. Mier, M. A. Andrade-Navarro, A. Bateman, P. Jarnot, A.
717 Gruca, M. Grynberg, A. V. Kajava, V. J. Promponas, M. Anisimova, K. S. Jakobsen, D.
718 Linke, Tandem repeats lead to sequence assembly errors and impose multi-level
719 challenges for genome and protein databases. *Nucleic Acids Research* **47**, 10994–
720 11006 (2019).

721 20. R. H. Toczydlowski, L. Liggins, M. R. Gaither, T. J. Anderson, R. L. Barton, J. T. Berg, S.
722 G. Beskid, B. Davis, A. Delgado, E. Farrell, M. Ghoojaei, N. Himmelsbach, A. E.
723 Holmes, S. R. Queeno, T. Trinh, C. A. Weyand, G. S. Bradburd, C. Riginos, R. J. Toonen,
724 E. D. Crandall, Poor data stewardship will hinder global genetic diversity surveillance.
725 *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2107934118 (2021).

726 21. E. D. Crandall, R. H. Toczydlowski, L. Liggins, A. E. Holmes, M. Ghoojaei, M. R.
727 Gaither, B. E. Wham, A. L. Pritt, C. Noble, T. J. Anderson, R. L. Barton, J. T. Berg, S. G.
728 Beskid, A. Delgado, E. Farrell, N. Himmelsbach, S. R. Queeno, T. Trinh, C. Weyand, A.
729 Bentley, J. Deck, C. Riginos, G. S. Bradburd, R. J. Toonen, Importance of timely
730 metadata curation to the global surveillance of genetic diversity. *Conservation*
731 *Biology* **37**, e14061 (2023).

732 22. Y.-N. Huang, P. V. Jaiswal, A. Rajes, A. Yadav, D. Yu, F. Liu, G. Scheg, E. Shih, G.
733 Boldirev, I. Nakashidze, A. Sarkar, J. H. Mehta, K. Wang, K. K. Patel, M. A. B. Mirza, K.
734 C. Hapani, Q. Peng, R. Ayyala, R. Guo, S. Kapur, T. Ramesh, D. Ciorbă, V. Munteanu, V.
735 Bostan, M. Dimian, M. S. Abedalthagafi, S. Mangul, The systematic assessment of
736 completeness of public metadata accompanying omics studies in the Gene
737 Expression Omnibus data repository. *Genome Biol* **26**, 274 (2025).

738 23. H. Bagheri, A. J. Severin, H. Rajan, Detecting and correcting misclassified sequences
739 in the large-scale public databases. *Bioinformatics* **36**, 4699–4705 (2020).

740 24. A. Gihawi, Y. Ge, J. Lu, D. Puiu, A. Xu, C. S. Cooper, D. S. Brewer, M. Pertea, S. L.
741 Salzberg, Major data analysis errors invalidate cancer microbiome findings. *mBio* **14**,
742 e01607-23 (2023).

743 25. V. Lupo, M. Van Vlierberghe, H. Vanderschuren, F. Kerff, D. Baurain, L. Cornet,
744 Contamination in Reference Sequence Databases: Time for Divide-and-Rule Tactics.
745 *Front. Microbiol.* **12**, 755101 (2021).

746 26. M. Steinegger, S. L. Salzberg, Terminating contamination: large-scale search
747 identifies more than 2,000,000 contaminated entries in GenBank. *Genome Biol* **21**,
748 115 (2020).

27. J. A. Byrne, N. Grima, A. Capes-Davis, C. Labbé, The Possibility of Systematic Research Fraud Targeting Under-Studied Human Genes: Causes, Consequences, and Potential Solutions. *Biomark Insights* **14**, 117727191982916 (2019).
28. R. A. K. Richardson, S. S. Hong, J. A. Byrne, T. Stoeger, L. A. N. Amaral, The entities enabling scientific fraud at scale are large, resilient, and growing rapidly. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2420092122 (2025).
29. Y.-M. Kim, J.-B. Poline, G. Dumas, Experimenting with reproducibility: a case study of robustness in bioinformatics. *GigaScience* **7** (2018).
30. J. Leipzig, D. Nüst, C. T. Hoyt, S. Soiland-Reyes, K. Ram, J. Greenberg, The role of metadata in reproducible computational research. *arXiv:2006.08589 [cs]* (2021).
31. F. Li, J. Hu, K. Xie, T.-C. He, Authentication of experimental materials: A remedy for the reproducibility crisis? *Genes Dis* **2**, 283 (2015).
32. N. S. Locatelli, P. B. McIntyre, N. O. Therkildsen, D. S. Baetscher, GenBank's reliability is uncertain for biodiversity researchers seeking species-level assignment for eDNA. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 32211–32212 (2020).
33. B. Tang, X. Hu, M. Yue, Contaminated bacterial genome data in the public domains: Evidence and solution. *Journal of Infection* **89**, 106369 (2024).
34. S. D. Chorlton, Ten common issues with reference sequence databases and how to mitigate them. *Front. Bioinform.* **4**, 1278228 (2024).
35. C.-T. Berezin, S. Peccoud, D. M. Kar, J. Peccoud, Cryptographic approaches to authenticating synthetic DNA sequences. *Trends in Biotechnology* **42**, 1002–1016 (2024).
36. C. J. Carlson, M. Granados, A. Phelan, N. Ramakrishnan, T. Poisot, The LISTEN principles for genetic sequence data governance and database engineering. *Nat Genet* **57**, 2099–2105 (2025).
37. D. Bloomfield, J. R. M. Black, O. Crook, N. Brandes, M. S. Hanke, T. V. Inglesby, A. Cicero, R. Pollack, T. Hernandez-Boussard, M. J. Imperiale, R. B. Altman, J. Pannu, Biological data governance in an age of AI. *Science* **391**, 558–561 (2026).
38. D. A. Yarmosh, J. G. Lopera, N. P. Puthuveetil, P. F. Combs, A. L. Reese, C. Tabron, A. E. Pierola, J. Duncan, S. R. Greenfield, R. Marlow, S. King, M. A. Riojas, J. Bagnoli, B. Benton, J. L. Jacobs, Comparative Analysis and Data Provenance for 1,113 Bacterial Genome Assemblies. *mSphere*, e00077-22 (2022).
39. R. R. Cheng, R. L. Bradford, A foundation for tomorrow's discoveries in cell biology. *Nat Cell Biol*, doi: 10.1038/s41556-025-01824-5 (2025).

- 783 40. Committee on Biological Collections: Their Past, Present, and Future Contributions
784 and Options for Sustaining Them, Board on Life Sciences, Division on Earth and Life
785 Studies, National Academies of Sciences, Engineering, and Medicine, *Biological*
786 *Collections: Ensuring Critical Research and Education for the 21st Century* (National
787 Academies Press, Washington, D.C., 2020;
788 <https://www.nationalacademies.org/publications/25592>).
- 789 41. C. L. Harmon, L. Castlebury, K. Boundy-Mills, K. Broders, A. M. Hyten, J. L. Jacobs, V.
790 K. Knight-Cannoni, D. Mollov, M. A. Riojas, P. Sharma, Standards of Diagnostic
791 Validation: Recommendations for reference collections. *PhytoFrontiers*TM, PHYTOFR-
792 05-22-0050-FI (2022).
- 793 42. F. Pleijel, U. Jondelius, E. Norlinder, A. Nygren, B. Oxelman, C. Schander, P. Sundberg,
794 M. Thollesson, Phylogenies without roots? A plea for the use of vouchers in
795 molecular phylogenetic studies. *Molecular Phylogenetics and Evolution* **48**, 369–371
796 (2008).
- 797 43. C. W. Thompson, K. L. Phelps, M. W. Allard, J. A. Cook, J. L. Dunnum, A. W. Ferguson,
798 M. Gelang, F. A. A. Khan, D. L. Paul, D. M. Reeder, N. B. Simmons, M. P. M. Vanhove, P.
799 W. Webala, M. Weksler, C. W. Kilpatrick, Preserve a Voucher Specimen! The Critical
800 Need for Integrating Natural History Collections in Infectious Disease Studies. *mBio*
801 **12**, e02698-20 (2021).
- 802 44. L. Kim, A. Lavrinienko, Z. Sebechlebska, S. Stoltenberg, N. A. Bokulich, Tier-based
803 standards for FAIR sequence data and metadata sharing in microbiome research.
804 *Nucleic Acids Res* **53**, gkaf777 (2025).
- 805 45. SRA Metadata and Submission Overview.
806 <https://www.ncbi.nlm.nih.gov/sra/docs/submitmeta/>.
- 807 46. J. Alnasir, H. P. Shanahan, Investigation into the annotation of protocol sequencing
808 steps in the sequence read archive. *Gigascience* **4**, s13742-015-0064–7 (2015).
- 809 47. M. N. Bernstein, A. Doan, C. N. Dewey, MetaSRA: normalized human sample-specific
810 metadata for the Sequence Read Archive. *Bioinformatics* **33**, 2914–2923 (2017).
- 811 48. R. S. Gonçalves, M. A. Musen, The variable quality of metadata about biological
812 samples used in biomedical experiments. *Sci Data* **6**, 190021 (2019).
- 813 49. S. Kanwal, F. Z. Khan, A. Lonie, R. O. Sinnott, Investigating reproducibility and
814 tracking provenance – A genomic workflow case study. *BMC Bioinformatics* **18**, 337
815 (2017).
- 816 50. J. A. Patel, D. A. Dean, C. H. King, N. Xiao, S. Koc, E. Minina, A. Golikov, P. Brooks, R.
817 Kahsay, R. Navelkar, M. Ray, D. Roberson, C. Armstrong, R. Mazumder, J. Keeney,

818 Bioinformatics tools developed to support BioCompute Objects. *Database (Oxford)*
819 **2021**, baab008 (2021).

820 51. K. Gierend, F. Krüger, S. Genehr, F. Hartmann, F. Siegel, D. Waltemath, T. Ganslandt,
821 A. A. Zeleke, Provenance Information for Biomedical Data and Workflows: Scoping
822 Review. *J Med Internet Res* **26**, e51297 (2024).

823 52. D. Fanelli, How Many Scientists Fabricate and Falsify Research? A Systematic Review
824 and Meta-Analysis of Survey Data. *PLoS ONE* **4**, e5738 (2009).

825 53. G. Gopalakrishna, G. Ter Riet, G. Vink, I. Stoop, J. M. Wicherts, L. M. Bouter,
826 Prevalence of questionable research practices, research misconduct and their
827 potential explanatory factors: A survey among academic researchers in The
828 Netherlands. *PLoS ONE* **17**, e0263023 (2022).

829 54. B. A. Sabel, E. Knaack, G. Gigerenzer, M.-I. Bilc, Fake publications in biomedical
830 science: red-flagging method indicates mass production. *Naunyn-Schmiedeberg's*
831 *Arch Pharmacol* **399**, 2943–2955 (2026).

832 55. B. Scancar, J. A. Byrne, D. Causeur, A. G. Barnett, Machine learning based screening
833 of potential paper mill publications in cancer research: methodological and cross
834 sectional study. *BMJ* **392**, e087581 (2026).

835 56. D. C. Richter, F. Ott, A. F. Auch, R. Schmid, D. H. Huson, MetaSim—A Sequencing
836 Simulator for Genomics and Metagenomics. *PLoS ONE* **3**, e3373 (2008).

837 57. Y. Ono, K. Asai, M. Hamada, PBSIM2: a simulator for long-read sequencers with a
838 novel generative model of quality scores. *Bioinformatics* **37**, 589–595 (2021).

839 58. Y. Li, S. Wang, C. Bi, Z. Qiu, M. Li, X. Gao, DeepSimulator1.5: a more powerful,
840 quicker and lighter simulator for Nanopore sequencing. *Bioinformatics* **36**, 2578–
841 2580 (2020).

842 59. H. Gourelé, O. Karlsson-Lindsjö, J. Hayer, E. Bongcam-Rudloff, Simulating Illumina
843 metagenomic data with InSilicoSeq. *Bioinformatics* **35**, 521–522 (2019).

844 60. A. Rambaut, N. C. Grass, Seq-Gen: an application for the Monte Carlo simulation of
845 DNA sequence evolution along phylogenetic trees. *Bioinformatics* **13**, 235–238
846 (1997).

847 61. G. Myers, A dataset generator for whole genome shotgun sequencing. *Proc Int Conf*
848 *Intell Syst Mol Biol*, 202–210 (1999).

849 62. E. Marshall, Fraud Strikes Top Genome Lab: Francis Collins, head of NIH's Human
850 Genome Project, has informed colleagues that a junior researcher in his lab faked
851 data in five papers Collins co-authored. *Science* **274**, 908–910 (1996).

- 852 63. G. Sangster, J. A. Luksenburg, Sharp Increase of Problematic Mitogenomes of Birds:
853 Causes, Consequences, and Remedies. *Genome Biology and Evolution* **13**, evab210
854 (2021).
- 855 64. M. P. Van Den Burg, D. R. Vieites, Bird genetic databases need improved curation and
856 error reporting to NCBI. *Ibis* **165**, 472–481 (2023).
- 857 65. A. Potti, H. K. Dressman, A. Bild, R. F. Riedel, G. Chan, R. Sayer, J. Cragun, H. Cottrill,
858 M. J. Kelley, R. Petersen, D. Harpole, J. Marks, A. Berchuck, G. S. Ginsburg, P. Febbo, J.
859 Lancaster, J. R. Nevins, Retraction Note: Genomic signatures to guide the use of
860 chemotherapeutics. *Nat Med* **17**, 135–135 (2011).
- 861 66. K. A. Baggerly, K. R. Coombes, Deriving chemosensitivity from cell lines: forensic
862 bioinformatics and reproducible research in high-throughput biology. *The Annals of*
863 *Applied Statistics*, 1309–1334 (2009).
- 864 67. NOT-OD-14-098: Findings of Research Misconduct, *Findings of Research Misconduct*
865 (2014). <https://grants.nih.gov/grants/guide/notice-files/NOT-OD-14-098.html>.
- 866 68. Retraction. *Endocrinology* **150**, 2976–2976 (2009).
- 867 69. M. S. Bradshaw, S. H. Payne, Detecting fabrication in large-scale molecular omics
868 data. *PLoS ONE* **16**, e0260395 (2021).
- 869 70. M. R. Bouadjene, K. Verspoor, J. Zobel, Automated detection of records in biological
870 sequence databases that are inconsistent with the literature. *Journal of Biomedical*
871 *Informatics* **71**, 229–240 (2017).
- 872 71. M. Sprang, J. Möllmann, M. A. Andrade-Navarro, J.-F. Fontaine, Overlooked poor-
873 quality patient samples in sequencing data impair reproducibility of published
874 clinically relevant datasets. *Genome Biol* **25**, 222 (2024).
- 875 72. Committee on Reproducibility and Replicability in Science, Board on Behavioral,
876 Cognitive, and Sensory Sciences, Committee on National Statistics, Division of
877 Behavioral and Social Sciences and Education, Nuclear and Radiation Studies
878 Board, Division on Earth and Life Studies, Board on Mathematical Sciences and
879 Analytics, Committee on Applied and Theoretical Statistics, Division on Engineering
880 and Physical Sciences, Board on Research Data and Information, Committee on
881 Science, Engineering, Medicine, and Public Policy, Policy and Global Affairs, National
882 Academies of Sciences, Engineering, and Medicine, *Reproducibility and Replicability*
883 *in Science* (National Academies Press, Washington, D.C., 2019;
884 <https://www.nationalacademies.org/publications/25303>).
- 885 73. P. I. Baykal, P. P. Łabaj, F. Markowetz, L. M. Schriml, D. J. Stekhoven, S. Mangul, N.
886 Beerewinkel, Genomic reproducibility in the bioinformatics era. *Genome Biol* **25**,
887 213 (2024).

- 888 74. R. S. Gonçalves, M. J. O'Connor, M. Martínez-Romero, J. Graybeal, M. A. Musen,
889 Metadata in the BioSample Online Repository are Impaired by Numerous Anomalies.
890 arXiv arXiv:1708.01286 [Preprint] (2017). <https://doi.org/10.48550/arXiv.1708.01286>.
- 891 75. A. Bernasconi, Data quality-aware genomic data integration. *Computer Methods and*
892 *Programs in Biomedicine Update* **1**, 100009 (2021).
- 893 76. J. T. Leek, R. D. Peng, Reproducible research can still be wrong: Adopting a prevention
894 approach. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 1645–1646 (2015).
- 895 77. B. Goudey, N. Geard, K. Verspoor, J. Zobel, Propagation, detection and correction of
896 errors using the sequence database network. *Briefings in Bioinformatics* **23**, bbac416
897 (2022).
- 898 78. S. V. Nguyen, V. H. Escobar, S. S. Ali, N. P. Puthuveetil, J. R. Petrone, J. L. Kirkland, K.
899 Gaffney, C. L. Tabron, N. Wax, J. Duncan, S. King, R. Marlow, A. L. Reese, D. A.
900 Yarmosh, H. H. McConnell, A. S. Fernandes, J. Bagnoli, B. Benton, J. L. Jacobs,
901 Reclassification of atypical *Moraxella catarrhalis* ATCC 23246 as *Moraxella veridica*
902 sp. nov. *International Journal of Systematic and Evolutionary Microbiology* **75** (2025).
- 903 79. N. Bouras, G. Dif, S. Belghit, S. V. Nguyen, J. L. Jacobs, O. Toumatia, S. S. Ebada, I.
904 Nouioui, Genome-based reclassification of the genus *Thermoanaerobacter*:
905 taxonomic emendations and new combinations. *Anaerobe* **97**, 103027 (2026).
- 906 80. S. V. Nguyen, D. Edwards, E. L. Vaughn, V. Escobar, S. Ali, J. H. Doss, J. T. Steyer, S.
907 Scott, W. Bchara, N. Bruns, E. Zelaya, A. Tran, D. Payne, J. R. Hauser, Expanding the
908 *Stenotrophomonas maltophilia* complex: phylogenomic insights, proposal of
909 *Stenotrophomonas forensis* sp. nov. and reclassification of two *Pseudomonas*
910 species. *International Journal of Systematic and Evolutionary Microbiology* **74** (2024).
- 911 81. L. P. Freedman, I. M. Cockburn, T. S. Simcoe, The Economics of Reproducibility in
912 Preclinical Research. *PLoS Biol* **13**, e1002165 (2015).
- 913 82. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K.
914 Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A.
915 Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back,
916 S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T.
917 Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P.
918 Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold.
919 *Nature* **596**, 583–589 (2021).
- 920 83. S. K. Burley, C. Bhikadiya, C. Bi, S. Bittrich, H. Chao, L. Chen, P. A. Craig, G. V.
921 Crichlow, K. Dalenberg, J. M. Duarte, S. Dutta, M. Fayazi, Z. Feng, J. W. Flatt, S. J.
922 Ganesan, S. Ghosh, D. S. Goodsell, R. K. Green, V. Guranovic, J. Henry, B. P. Hudson,
923 I. Khokhriakov, C. L. Lawson, Y. Liang, R. Lowe, E. Peisach, I. Persikova, D. W. Piehl, Y.

- 924 Rose, A. Sali, J. Segura, M. Sekharan, C. Shao, B. Vallat, M. Voigt, B. Webb, J. D.
925 Westbrook, S. Whetstone, J. Y. Young, A. Zalevsky, C. Zardecki, RCSB Protein Data
926 bank: Tools for visualizing and understanding biological macromolecules in 3D.
927 *Protein Sci* **31**, e4482 (2022).
- 928 84. J. Y. Young, J. D. Westbrook, Z. Feng, E. Peisach, I. Persikova, R. Sala, S. Sen, J. M.
929 Berrisford, G. J. Swaminathan, T. J. Oldfield, A. Gutmanas, R. Igarashi, D. R.
930 Armstrong, K. Baskaran, L. Chen, M. Chen, A. R. Clark, L. Di Costanzo, D.
931 Dimitropoulos, G. Gao, S. Ghosh, S. Gore, V. Guranovic, P. M. S. Hendrickx, B. P.
932 Hudson, Y. Ikegawa, Y. Kengaku, C. L. Lawson, Y. Liang, L. Mak, A. Mukhopadhyay, B.
933 Narayanan, K. Nishiyama, A. Patwardhan, G. Sahni, E. Sanz-García, J. Sato, M. R.
934 Sekharan, C. Shao, O. S. Smart, L. Tan, G. Van Ginkel, H. Yang, M. A. Zhuravleva, J. L.
935 Markley, H. Nakamura, G. Kurisu, G. J. Kleywegt, S. Velankar, H. M. Berman, S. K.
936 Burley, Worldwide Protein Data Bank biocuration supporting open access to high-
937 quality 3D structural biology data. *Database* **2018** (2018).
- 938 85. G. Brixi, M. G. Durrant, J. Ku, M. Naghipourfar, M. Poli, G. Sun, G. Brockman, D.
939 Chang, A. Fanton, G. A. Gonzalez, S. H. King, D. B. Li, A. T. Merchant, E. Nguyen, C.
940 Ricci-Tam, D. W. Romero, J. C. Schmok, A. Taghibakhshi, A. Vorontsov, B. Yang, M.
941 Deng, L. Gorton, N. Nguyen, N. K. Wang, M. T. Pearce, E. Simon, E. Adams, Z. J.
942 Amador, E. A. Ashley, S. A. Baccus, H. Dai, S. Dillmann, S. Ermon, D. Guo, M. H.
943 Herschl, R. Ilango, K. Janik, A. X. Lu, R. Mehta, M. R. K. Mofrad, M. Y. Ng, J. Pannu, C.
944 Ré, J. St. John, J. Sullivan, J. Tey, B. Viggiano, K. Zhu, G. Zynda, D. Balsam, P. Collison,
945 A. B. Costa, T. Hernandez-Boussard, E. Ho, M.-Y. Liu, T. McGrath, K. Powell, S. Pinglay,
946 D. P. Burke, H. Goodarzi, P. D. Hsu, B. L. Hie, Genome modelling and design across all
947 domains of life with Evo 2. *Nature*, 1–13 (2026).
- 948 86. D. H. Parks, P.-A. Chaumeil, A. J. Mussig, C. Rinke, M. Chuvochina, P. Hugenholtz,
949 GTDB release 10: a complete and systematic taxonomy for 715 230 bacterial and 17
950 245 archaeal genomes. *Nucleic Acids Research* **54**, D743–D754 (2026).
- 951 87. D. Paez-Espino, I.-M. A. Chen, K. Palaniappan, A. Ratner, K. Chu, E. Szeto, M. Pillay, J.
952 Huang, V. M. Markowitz, T. Nielsen, M. Huntemann, T. B. K. Reddy, G. A. Pavlopoulos,
953 M. B. Sullivan, B. J. Campbell, F. Chen, K. McMahon, S. J. Hallam, V. Deneff, R.
954 Cavicchioli, S. M. Caffrey, W. R. Streit, J. Webster, K. M. Handley, G. H. Salekdeh, N.
955 Tsesmetzis, J. C. Setubal, P. B. Pope, W.-T. Liu, A. R. Rivers, N. N. Ivanova, N. C.
956 Kyrpides, IMG/VR: a database of cultured and uncultured DNA Viruses and
957 retroviruses. *Nucleic Acids Research* **45**, gkw1030 (2016).
- 958 88. A. P. Camargo, L. Call, S. Roux, S. Nayfach, M. Huntemann, K. Palaniappan, A.
959 Ratner, K. Chu, S. Mukherjee, T. B. K. Reddy, I.-M. A. Chen, N. N. Ivanova, E. A. Elie-
960 Fadrosch, T. Woyke, D. A. Baltrus, S. Castañeda-Barba, F. de la Cruz, B. E. Funnell, J. P.
961 J. Hall, A. Mukhopadhyay, E. P. C. Rocha, T. Stalder, E. Top, N. C. Kyrpides, IMG/PR: a
962 database of plasmids from genomes and metagenomes with rich annotations and
963 metadata. *Nucleic Acids Research* **52**, D164–D173 (2024).

- 964 89. H. Dalla-Torre, L. Gonzalez, J. Mendoza-Revilla, N. Lopez Carranza, A. H.
965 Grzywaczewski, F. Oteri, C. Dallago, E. Trop, B. P. de Almeida, H. Sirelkhatim, G.
966 Richard, M. Skwark, K. Beguir, M. Lopez, T. Pierrot, Nucleotide Transformer: building
967 and evaluating robust foundation models for human genomics. *Nat Methods* **22**,
968 287–297 (2025).
- 969 90. Y. Zeng, J. Xie, N. Shangguan, Z. Wei, W. Li, Y. Su, S. Yang, C. Zhang, J. Zhang, N. Fang,
970 H. Zhang, Y. Lu, H. Zhao, J. Fan, W. Yu, Y. Yang, CellFM: a large-scale foundation
971 model pre-trained on transcriptomics of 100 million human cells. *Nat Commun* **16**,
972 4679 (2025).
- 973 91. G. Heimberg, T. Kuo, D. J. DePianto, O. Salem, T. Heigl, N. Diamant, G. Scalia, T.
974 Biancalani, S. J. Turley, J. R. Rock, H. Corrada Bravo, J. Kaminker, J. A. Vander Heiden,
975 A. Regev, A cell atlas foundation model for scalable search of similar human cells.
976 *Nature* **638**, 1085–1094 (2025).
- 977 92. M. Hemberg, F. Marini, S. Ghazanfar, A. Al Ajami, N. Abassi, B. Anchang, B. A.
978 Benayoun, Y. Cao, K. Chen, Y. Cuesta-Astro, Z. DeBruine, C. A. Dendrou, I. De
979 Vlaminck, K. Imkeller, I. Korsunsky, A. R. Lederer, J. J. Li, P. Meysman, C. L. Miller, K. A.
980 Mullan, U. Ohler, P. Panwar, N. Patikas, J. Schuck, J. H. Y. Siu, T. J. Triche, A. Tsankov,
981 S. W. van der Laan, M. Yajima, J. Yang, F. Zanini, I. Jelic, Insights, opportunities, and
982 challenges provided by large cell atlases. *Genome Biol* **26**, 358 (2025).
- 983 93. H. Feng, L. Wu, B. Zhao, C. Huff, J. Zhang, J. Wu, L. Lin, P. Wei, C. Wu, Benchmarking
984 DNA foundation models for genomic and genetic tasks. *Nat Commun* **16**, 10780
985 (2025).
- 986 94. R. Joeres, D. B. Blumenthal, O. V. Kalinina, Data splitting to avoid information leakage
987 with DataSAIL. *Nat Commun* **16**, 3337 (2025).
- 988 95. S. V. Nguyen, N. P. Puthuveetil, J. R. Petrone, J. L. Kirkland, K. Gaffney, C. L. Tabron, N.
989 Wax, J. Duncan, S. King, R. Marlow, A. L. Reese, D. A. Yarmosh, H. H. McConnell, A. S.
990 Fernandes, J. Bagnoli, B. Benton, J. L. Jacobs, The ATCC genome portal: 3,938
991 authenticated microbial reference genomes. *Microbiol Resour Announc* **13**, e01045-
992 23 (2024).
- 993 96. J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J.
994 Wilson, J. Lehár, G. V. Kryukov, D. Sonkin, A. Reddy, M. Liu, L. Murray, M. F. Berger, J. E.
995 Monahan, P. Morais, J. Meltzer, A. Korejwa, J. Jané-Valbuena, F. A. Mapa, J. Thibault, E.
996 Bric-Furlong, P. Raman, A. Shipway, I. H. Engels, J. Cheng, G. K. Yu, J. Yu, P. Aspesi, M.
997 De Silva, K. Jagtap, M. D. Jones, L. Wang, C. Hatton, E. Palescandolo, S. Gupta, S.
998 Mahan, C. Sougnez, R. C. Onofrio, T. Liefeld, L. MacConaill, W. Winckler, M. Reich, N.
999 Li, J. P. Mesirov, S. B. Gabriel, G. Getz, K. Ardlie, V. Chan, V. E. Myer, B. L. Weber, J.
1000 Porter, M. Warmuth, P. Finan, J. L. Harris, M. Meyerson, T. R. Golub, M. P. Morrissey,

1001 W. R. Sellers, R. Schlegel, L. A. Garraway, The Cancer Cell Line Encyclopedia enables
1002 predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).

1003 97. B. Haibe-Kains, N. El-Hachem, N. J. Birkbak, A. C. Jin, A. H. Beck, H. J. W. L. Aerts, J.
1004 Quackenbush, Inconsistency in large pharmacogenomic studies. *Nature* **504**, 389–
1005 393 (2013).

1006 98. M. Bouhaddou, M. S. DiStefano, E. A. Riesel, E. Carrasco, H. Y. Holzapfel, D. C. Jones,
1007 G. R. Smith, A. D. Stern, S. S. Somani, T. V. Thompson, M. R. Birtwistle, Drug response
1008 consistency in CCLE and CGP. *Nature* **540**, E9–E10 (2016).

1009 99. J. P. Mpindi, B. Yadav, P. Östling, P. Gautam, D. Malani, A. Murumägi, A. Hirasawa, S.
1010 Kangaspeska, K. Wennerberg, O. Kallioniemi, T. Aittokallio, Consistency in drug
1011 response profiling. *Nature* **540**, E5–E6 (2016).

1012 100. U. Ben-David, B. Siranosian, G. Ha, H. Tang, Y. Oren, K. Hinohara, C. A. Strathdee, J.
1013 Dempster, N. J. Lyons, R. Burns, A. Nag, G. Kugener, B. Cimini, P. Tsvetkov, Y. E.
1014 Maruvka, R. O’Rourke, A. Garrity, A. A. Tubelli, P. Bandopadhyay, A. Tsherniak, F.
1015 Vazquez, B. Wong, C. Birger, M. Ghandi, A. R. Thorner, J. A. Bittker, M. Meyerson, G.
1016 Getz, R. Beroukhi, T. R. Golub, Genetic and transcriptional evolution alters cancer
1017 cell line drug response. *Nature* **560**, 325–330 (2018).

1018 101. R. E. Hynds, E. Vladimirov, Sam. M. Janes, The secret lives of cancer cell lines.
1019 *Disease Models & Mechanisms* **11**, dmm037366 (2018).

1020 102. H. Sichtig, T. Minogue, Y. Yan, C. Stefan, A. Hall, L. Tallon, L. Sadzewicz, S. Nadendla,
1021 W. Klimke, E. Hatcher, M. Shumway, D. L. Aldea, J. Allen, J. Koehler, T. Slezak, S.
1022 Lovell, R. Schoepp, U. Scherf, FDA-ARGOS is a database with public quality-
1023 controlled reference genomes for diagnostic use and regulatory science. *Nat*
1024 *Commun* **10**, 3313 (2019).

1025 103. J. Dicks, M.-A. Fazal, K. Oliver, N. E. Grayson, J. D. Turnbull, E. Bane, E. Burnett, A.
1026 Deheer-Graham, N. Holroyd, D. Kaushal, J. Keane, G. Langridge, J. Lomax, H.
1027 McGregor, S. Picton, M. Quail, D. Singh, A. Tracey, J. Korlach, J. E. Russell, S.
1028 Alexander, J. Parkhill, NCTC3000: a century of bacterial strain collecting leads to a
1029 rich genomic data resource. *Microbial Genomics* **9** (2023).

1030 104. S. Kato, S. Masuda, A. Shibata, T. Itoh, M. Sakamoto, K. Shirasu, M. Ohkuma, Whole-
1031 genome sequencing of diverse 351 cultured prokaryotes including as-yet-
1032 unsequenced type strains. *Genome Res* **36**, 875–884 (2026).

1033 105. L. Wu, K. McCluskey, P. Desmeth, S. Liu, S. Hideaki, Y. Yin, O. Moriya, T. Itoh, C. Y.
1034 Kim, J.-S. Lee, Y. Zhou, H. Kawasaki, M. H. Hazbón, V. Robert, T. Boekhout, N. Lima, L.
1035 Evtushenko, K. Boundy-Mills, B. Bunk, E. R. B. Moore, L. Eurwilaichitr, S. Ingsriswang,
1036 H. Shah, S. Yao, T. Jin, J. Huang, W. Shi, Q. Sun, G. Fan, W. Li, X. Li, I. Kurtböke, J. Ma,

- 1037 The global catalogue of microorganisms 10K type strain sequencing project: closing
1038 the genomic gaps for the validly published prokaryotic and fungi species.
1039 *GigaScience* **7** (2018).
- 1040 106. P. S. G. Chain, D. V. Grafham, R. S. Fulton, M. G. FitzGerald, J. Hostetler, D. Muzny, J.
1041 Ali, B. Birren, D. C. Bruce, C. Buhay, J. R. Cole, Y. Ding, S. Dugan, D. Field, G. M.
1042 Garrity, R. Gibbs, T. Graves, C. S. Han, S. H. Harrison, S. Highlander, P. Hugenholtz, H.
1043 M. Khouri, C. D. Kodira, E. Kolker, N. C. Kyrpides, D. Lang, A. Lapidus, S. A. Malfatti, V.
1044 Markowitz, T. Metha, K. E. Nelson, J. Parkhill, S. Pitluck, X. Qin, T. D. Read, J. Schmutz,
1045 S. Sozhamannan, P. Sterk, R. L. Strausberg, G. Sutton, N. R. Thomson, J. M. Tiedje, G.
1046 Weinstock, A. Wollam, Genomic Standards Consortium Human Microbiome Project
1047 Jumpstart Consortium, J. C. Detter, Genome Project Standards in a New Era of
1048 Sequencing. *Science* **326**, 236–237 (2009).
- 1049 107. J. Wieczorek, D. Bloom, R. Guralnick, S. Blum, M. Döring, R. Giovanni, T. Robertson,
1050 D. Viegla, Darwin Core: An Evolving Community-Developed Biodiversity Data
1051 Standard. *PLoS ONE* **7**, e29715 (2012).
- 1052 108. G. Droege, K. Barker, O. Seberg, J. Coddington, E. Benson, W. G. Berendsohn, B.
1053 Bunk, C. Butler, E. M. Cawsey, J. Deck, M. Döring, P. Flemons, B. Gemeinholzer, A.
1054 Güntsch, T. Hollowell, P. Kelbert, I. Kostadinov, R. Kottmann, R. T. Lawlor, C. Lyal, J.
1055 Mackenzie-Dodds, C. Meyer, D. Mulcahy, S. Y. Nussbeck, É. O'Tuama, T. Orrell, G.
1056 Petersen, T. Robertson, C. Söhngen, J. Whitacre, J. Wieczorek, P. Yilmaz, H. Zetsche,
1057 Y. Zhang, X. Zhou, The Global Genome Biodiversity Network (GGBN) Data Standard
1058 specification. *Database* **2016**, baw125 (2016).
- 1059 109. E. J. Griffiths, R. E. Timme, A. J. Page, N.-F. Alikhan, D. Fornika, F. Maguire, C. I.
1060 Mendes, S. H. Tausch, A. Black, T. R. Connor, G. H. Tyson, D. M. Aanensen, B. Alcock,
1061 J. Campos, A. Christoffels, A. Gonçalves Da Silva, E. Hodcroft, W. W. L. Hsiao, L. S.
1062 Katz, S. M. Nicholls, P. E. Oluniyi, I. B. Olawoye, A. R. Raphenya, A. T. R. Vasconcelos,
1063 A. A. Witney, D. R. MacCannell, The PHA4GE SARS-CoV-2 Contextual Data
1064 Specification for Open Genomic Epidemiology. *Biology and Life Sciences* [Preprint]
1065 (2020). <https://doi.org/10.20944/preprints202008.0220.v1>.
- 1066 110. E. J. Griffiths, R. E. Timme, C. I. Mendes, A. J. Page, N.-F. Alikhan, D. Fornika, F.
1067 Maguire, J. Campos, D. Park, I. B. Olawoye, P. E. Oluniyi, D. Anderson, A. Christoffels,
1068 A. G. da Silva, R. Cameron, D. Dooley, L. S. Katz, A. Black, I. Karsch-Mizrachi, T.
1069 Barrett, A. Johnston, T. R. Connor, S. M. Nicholls, A. A. Witney, G. H. Tyson, S. H.
1070 Tausch, A. R. Raphenya, B. Alcock, D. M. Aanensen, E. Hodcroft, W. W. L. Hsiao, A. T.
1071 R. Vasconcelos, D. R. MacCannell, Future-proofing and maximizing the utility of
1072 metadata: The PHA4GE SARS-CoV-2 contextual data specification package.
1073 *GigaScience* **11**, giac003 (2022).

- 1074 111. A. Klie, B. Y. Tsui, S. Mollah, D. Skola, M. Dow, C.-N. Hsu, H. Carter, Increasing
1075 metadata coverage of SRA BioSample entries using deep learning–based named
1076 entity recognition. *Database (Oxford)* **2021**, baab021 (2021).
- 1077 112. V. Simonyan, J. Goecks, R. Mazumder, Biocompute Objects—A Step towards
1078 Evaluation and Validation of Biomedical Scientific Computations. *PDA Journal of*
1079 *Pharmaceutical Science and Technology* **71**, 136–146 (2017).
- 1080 113. Food and Drug Administration, Electronic Submissions; Data Standards; Support for
1081 the International Institute of Electrical and Electronics Engineers Bioinformatics
1082 Computations and Analyses Standard for Bioinformatic Workflows (2020).
1083 <https://www.federalregister.gov/d/2020-15771>.
- 1084 114. P. W. Harrison, J. Fan, D. Richardson, L. Clarke, D. Zerbino, G. Cochrane, A. L.
1085 Archibald, C. J. Schmidt, P. Flicek, FAANG , establishing metadata standards,
1086 validation and best practices for the farmed and companion animal community.
1087 *Animal Genetics* **49**, 520–526 (2018).
- 1088 115. ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the
1089 human genome. *Nature* **489**, 57–74 (2012).

Acknowledgements

Sincere thanks is given to colleagues who reviewed this manuscript prior to publication, including Cara Wilder, David Yarmosh, Emily White, Scott Nguyen, David Molik, Sterling Sawaya, James Crill, Patrick Boyle, and Ruth Cheng.

Funding: This work was supported by internal funding from ATCC.

Author Contributions: The conceptualization, investigation, and writing of this manuscript was done by J.L.J.

Competing Interests: The author declares no competing interests.