Published online DD MM YYYY *Nucleic Acids Research, YYYY, Vol. xx, No. xx* ***1–12***
doi:10.1093/nar/gkn000

DATA RESOURCES AND ANALYSES
FAIR Header Reference genome: A TRUSTworthy standard

Adam Wright[1,*], Mark D Wilkinson[2], Chris Mungall[3], Scott Cain[1], Stephen Richards[4], Paul Sternberg[5], Ellen Provin[6], Jonathan L Jacobs[7], Scott Geib[8], Daniela Raciti[5], Karen Yook[5], Lincoln Stein[1], David C Molik[9, *]

[1] Adaptive Oncology Program, Ontario Institute for Cancer Research, 661 University Avenue Suite 500, Toronto, ON M5G 0A3, Canada [2] Departamento de Biotecnología-Biología Vegetal, Escuela Técnica Superior de Ingeniería Agronómica, Alimentaria y de Biosistemas,Centro de Biotecnología y Genómica de Plantas (CBGP, UPM-INIA/CSIC), Universidad Politécnica de Madrid (UPM) - Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA/CSIC), Pozuelo de Alarcón, Madrid, ES, Spain [3] Biosystems Data Science, Lawrence Berkeley National Laboratory, Building: 977, 1 Cyclotron Rd, Berkeley, CA 94720 USA [4] Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, MS: BCM226, Houston, TX 77030, USA [5] Division of Biology and Biological Engineering 140-18, California Institute of Technology, Pasadena, CA 91125, USA [6] Department of Horticultural Studies, Texas A&M University, HFSB 204, TAMU 2133, College Station, TX 77848, USA [7] American Type Culture Collection, 10801 University Blvd, Manassas, VA 20110, USA [8] Tropical Pest Genetics and Molecular Biology Research Unit, Daniel K. Inouye U.S. Pacific Basin Agricultural Research Center, United States Department of Agriculture, Agricultural Research Service, 64 Nowelo St, Hilo HI 96720 USA [9] Arthropod-borne Animal Diseases Research Unit, Center for Grain and Animal Health Research United States Department of Agriculture, Agricultural Research Service, 1515 College Ave, Manhattan, KS 66502 USA

Received YYYY-MM-DD; Revised YYYY-MM-DD; Accepted YYYY-MM-DD

ABSTRACT

The lack of interoperable data standards among reference genome data-sharing platforms inhibits cross-platform analysis while increasing the risk of data provenance loss. Here, we describe the FAIR-bioHeaders Reference genome (FHR), a metadata standard guided by the principles of Findability, Accessibility, Interoperability, and Reuse (FAIR) in addition to the principles of Transparency, Responsibility, User focus, Sustainability, and Technology (TRUST). The objective of FHR is to provide an extensive set of data serialisation methods and minimum data field requirements while still maintaining extensibility, flexibility, and expressivity in an increasingly decentralised genomic data ecosystem. The effort needed to implement FHR is low; FHR's design philosophy ensures easy implementation while retaining the benefits gained from recording both machine and human-readable provenance.

INTRODUCTION

Importance of reference genomes

The large and ever-increasing number of well-characterised reference genomes has become a prerequisite for many essential analyses, including cross-species comparisons and population genomics studies of model and non-model systems (1). Unfortunately, there are no file header standards for the metadata describing these essential resources, and even basic fields, such as ”species” and ”strain”, are missing from the common reference genome data standards. Instead, such metadata must be kept separately from the files containing the reference genome data, raising the risk of gaps in data provenance and human copying errors and imposing a burden on computational biologists and developers of analytic software alike (2, 3). The FAIR-bioHeaders Reference genome (FHR) specification aims to provide a standard to maintain the provenance of reference genomes that is translatable across storage and analysis platforms.

History of FASTA

The FASTA file format is widely used in genomic analysis as the repository of reference genome sequence information. FASTA was developed in 1985 (4) and is still widely used for reference genomes, but during that time the genomics ecosystem has changed dramatically; the first journal mention

*To whom correspondence should be addressed.
Tel: +1 (979) 260-9329, Fax: +1 (979) 260-9333
Email: adam.wright@oicr.on.ca, david.molik@usda.gov

of a centralised DNA repository was in Elke Jordan and Christine Carrico’s Science Letter in 1982 (5). The FASTA format specification (originally the ”Pearson format”) was created by William Pearson and David Lipman in 1985 (4), but has since been maintained by the National Center for Biotechnology Information (NCBI) at the U.S. National Institutes of Health (4).

It used to be that reference genome files resided in a small number of trusted repositories, such as GenBank, and were downloaded from there to the bioinformatician’s local computer system, it is increasingly the case that these files are modified, reannotated, and redistributed in a decentralised manner (6, 7, 8). This was not the consequence of any deliberative process. Instead, it occurred organically as a consequence of the shared ownership of the reference data and the collaborative nature of science (9). Decentralisation also occurs when multiple websites host genomic data instead of a single authority, and it is unlikely that this process will reverse (6). Decentralisation carries risks, not least of which is the loss of provenance metadata that may occur when the files are transferred among resources and when users download the genomes for local processing. Furthermore, loss of provenance reduces the ability of users to ensure that data are what they claim to be, potentially causing confusion, propagating errors in subsequent analysis, and increasing overall time and effort to reuse data (10).

Deficiencies of FASTA for reference genomes

It is highly problematic that reference genome FASTA files contain no intrinsic information that describes the nature and provenance of their contents; all provenance information must come from external sources and be linked to the file name or checksum. However, both of these methods are prone to loss of information. Files can be easily renamed or overwritten, and when the name has changed, the link to provenance information can be difficult to recover. Checksums, an algorithmically unique representation of a file that can be compared for accuracy, are also brittle. Commonly performed file manipulations, such as introducing carriage returns when a file generated on a Linux system is opened in a Windows text editor, introduce alterations that do not affect the semantics of the file, but completely change the checksum. By relying on external information for the provenance of the file, bioinformaticians risk associating incorrect metadata with the genome file or even being unable to locate the metadata at all (11, 12).

Differences can arise when a reference genome is replicated across platforms or devices (e.g. renaming of files or contigs, removal of contigs that fail to meet some criteria such as minimum length, the removal and addition of metadata, etc.) leading to a gradual divergence of reference genome files and their metadata (i.e., the genome data and metadata divergence problem, divergence problems are described by Haslhofer 2010 (13)). Furthermore, discrepancies can arise when a genome assembly is updated with additional data, due to inexact version matching from multiple genome assembly versions and user updates across platforms. To address the discrepancies that arise from replication, what could be called a reference genome authority is typically implemented (e.g. https://www.ncbi.nlm.nih.gov/assembly, https://www.ncbi.nlm.nih.gov/genbank/, https://www.ddbj.nig.ac.jp/index-e.html), a central site that provides the authoritative version of a reference genome and its origin. While reference genome authorities are the ideal solution, in the current biological data environment, a reference genome authority is not always a practical solution. Two recent examples of reference genomes being published in multiple locations illustrate not only the need for genome hosting to be available, but also the necessity of decentralised assembly hosting and how file-level discrepancies can be introduced.

One example is the American Type Culture Collection (ATCC), a major biorepository and living culture collection that provides researchers with the physical strains and cell lines needed for their research. Historically, materials obtained from ATCC have been subjected to whole genome sequencing by researchers using those materials in their own research. The resulting genome assemblies produced by researchers are often submitted to the NCBI Assembly reference genome database in order to disseminate and share the data (14). The NCBI Assembly reference database, in this case, may be thought of as the reference genome authority. However, gaps in genomics data quality, data provenance and the traceability of materials used by researchers have contributed significantly to the scientific reproducibility crisis (reviewed in Hirsch 2019: (15)). In response to issues of provenance and authenticity, ATCC launched the https://genomes.atcc.org/ to establish its own quality control and provenance standards associated with genome references that represent authentic ATCC materials (16). ATCC has thus far produced over 4,000 high-quality or closed reference genomes for microbes within the ATCC collection, all under an ISO 9000 controlled quality assurance framework. This presents some dilemmas, however, as the NCBI Assembly database includes (for example) genome references for bacterial strains that have serious gaps in metadata or include substantial errors in their genome assembly when compared to the ATCC Genome Portal reference (17). Reducing discrepancies between genome references for the ”same” organism can be aided by improving our ability to include crucial metadata about the origins of and means by which each genome reference is created in-line with the sequence data itself.

Another example of discrepencies that can arise from gaps in provenance can be found in molecular data portals and genome browsers. Several organism-focused genome data portals, such as AgBase (18), FlyBase (19), SoyBase (20), wFleaBase (21), WormBase (22), VectorBase (23), Ensembl (24), and others (25), publish annotations that are not found in the NCBI Assembly database. In some cases, these annotations and associated genomes cannot be submitted due to data ownership conflicts. These genome browsers and data repositories are often associated with a larger consortium that is working to answer questions of interest to the relevant scientific communities. Examples of such consortiums are the i5k (26, 27) Workspace (28), a collaborative effort to annotate arthropod genomes, and the Alliance of Genome Resources (The Alliance) (29) a centralised resource Model Organism resource. However, the data that these communities require have specific requirements, which can lead to the data portals and genome browsers becoming the primary source of their scientific communities’ reference genomes. Regardless of the resources that currently host the genome, there is no link to the source of the file, including its metadata, and associated publications. Since decentralisation is currently ongoing, it is unreasonable to imagine a world in which reference genome assemblies are not shared across platforms.

In this paper, we present the FHR FASTA header specification, which has been developed to address the genome data and metadata divergence problem. A key benefit of FHR is that it minimises the technical impact of adding provenance metadata to FASTA reference genome files by utilising legacy features of the file format instead of adding completely new ones. FHR is designed to enable FAIR and TRUST principles(30, 31), and to reduce the risk of data loss by ensuring that the provenance metadata is tied to the reference genome.

METHODS

FHR version one is publicly available on GitHub within the organisation FAIR-bioHeaders. There are two relevant repositories: the specification and FHR-related tools (i.e. the FHR conversion and validation toolkit). The specification is codified within JSON following the JSON Schema specification. The related tools are written in Python and use this JSON schema to validate FHR-specified files. To ensure that the software is as maintainable as possible, we chose to use a minimal set of well-established dependencies. The Python libraries on which FHR converter tools depend include: YAML, JSON, Microdata v0.8.0, re (REGEX), hashlib, and JSONSchema v4.17.3 libraries. The tool requires Python 3.6 or higher and can be installed using the Python package setuptools version 42 or higher. These dependencies are all that are required to validate and convert FHR files. The goal of using so few dependencies is to reduce the potential for issues that arise when installing the tools and reduce the effort required to maintain the tools.

RESULTS

The core constituent FHR fields were determined by considering the hypothetical, but unlikely, scenario of catastrophic loss of all copies of a reference genome. In this hypothetical scenario, all digital copies of the genome assembly have been lost, along with raw data. To reconstruct the genome, the following fields would be required: the location of the biological materials used to create the genome, the sequencing instruments used, and the assembly tools used to assemble and quality-check the genome. Therefore, FHR records the location of the biological materials, the sequencing instruments, and the assembly software tools. Furthermore, FHR records other information that would be useful in recovering such a genome: the metadata author of the FHR document, the assembler used, and other documentation either in the FHR instance or found in any scholarly articles associated with the genome. FHR also records funding and licencing information, related links, and the name and version of the genome. The intent of FHR is to strike a balance between only forcing users to provide the minimal information about a reference genome assembly but also be flexable enough to include other information. Therefore, the required fields are the absolute minimum to provide provenance of the data, and the optional fields focus on providing other useful information, and flexibility. It is recognized that more information on sample preparation and data processing could be added to the header but to keep the specification reasonably concise not all possible fields were added to the specification.

Required fields

The FHR specification has nine required fields: `schema`, `schemaVersion`, `genome`, `taxon`, `version`, `assemblyAuthor`, `metadataAuthor`, `dateCreated`, and `checksum`. Please refer to (Table 1, and Fig. 1).

The `schema` field indicates which JSON schema specification the metadata adheres to. In combination with the `schemaVersion` field, it allows users and software to know the exact format of the metadata to which the header conforms.

The `genome` field is a string that is used to refer to the common name of the genome. The field contents are chosen by those generating the reference genome and can be a human-readable name, an alphanumeric ID, or a URI. The latter options are designed to simplify automated analysis.

`taxon` allows the user to specify the species in both a human-readable string as well as a link to identifiers.org to provide more information on the genome.

The `assemblyAuthor` field is a list of authors who participated in the generating of the genome; FHR supports multiple assembly authors. Typically, these authors would be those who contributed to the original paper(s) describing the sequencing and annotation of the genome, but this choice is left to those who generated the reference genome. In contrast, `metadataAuthor` identifies the person who authored the header metadata; FHR supports multiple metadata authors, as multiple individuals or organisations can contribute to the recording of its provenance. By providing both authorship fields, FHR supports situations where the creators of the FHR metadata are not the same as the creators of the assembly itself; this is useful when FHR metadata have been added after the fact by another group.

The `dateCreated` field specifies the creation date for the complete reference genome file. It works hand in hand with the `checksum` field, which provides a way of confirming that neither the data nor the metadata have changed since the file was created. We consider the checksum field to be one of the FHR format’s most useful features. In addition to being used to ensure that the file has not been corrupted, the checksum can also be recorded by the pipeline to keep track of the exact assembly used within an analysis run. Pipelines typically record the name of the assembly (e.g. HG19), but it is not uncommon for various *ad hoc* variants of the assembly to be circulated within the community, causing ambiguity. The checksum uniquely identifies the assembly and its metadata, thereby providing an identifier that can be used by analytic pipelines to unambiguously declare which assembly their results are based on.

Together, these nine fields allow researchers to identify the format of the metadata, identify the contents of the file, and keep track of who made the file and when, thereby enabling provenance-tracking.

Optional fields

In addition to the nine required fields, FHR encourages the addition of up to 11 optional fields. These additional

Table 1. FHR Required Fields

Field	Example	Description
schema	https://raw.githubusercontent.com/FAIR-bioHeaders/FHR-Specification/main/fhr.json	URI to the FHR schema
scehmaVersion	1	Version of FHR
genome	Example species	Name of the genome
taxon		taxonomic identification of the genome, taxon itself is purely an organisational component to hold taxon name and taxon URI
taxon name	Example species	(property of taxon, name, or common name of taxon)
taxon uri	https://identifiers.org/taxonomy:0000	(property of taxon, URL of the taxon information, to a registry if possible)
version	2.3	Version number of genome
metadataAuthor		Author of the FHR Instance (Person or Organisation), not the genome, multiple metadataAuthors are allowed, metadataAuthor itself is purely an organisational component.
metadataAuthor name	John Doe	(property of author, the name of the author)
metadataAuthor uri	https://orcid.org/0000-0002-1983-4588	(property of author, the URL of the author)
assemblyAuthor		Assembler of the Genome (Person or Org), multiple assemblyAuthors are allowed, assemblyAuthor itself is purely an organisational component
assemblyAuthor name	Jane Doe	(property of assembler, the name of the assembler)
assemblyAuthor uri	https://orcid.org/0000-0002-9511-5139	(property of the assembler, the URL of the assembler)
dateCreated	2022-03-21	the date the genome assembly was created
checksum	md5:a3d5d9146c3992b7ed6724409ba28aa9	algorithm and hash for the checksum of reference genomes

Figure 1. Minimal FHR FASTA header example with sequence

```
;~schema: https://raw.githubusercontent.com/FAIR-bioHeaders/FHR-Specification/main/
   fhr.json
;~schemaVersion: 1.0
;~genome: Example species
;~taxon:
;~  name: Example species
;~  uri: https://identifiers.org/taxonomy:0000
;~version: 0.0.1
;~metadataAuthor:
;~  name: Adam Wright
;~  uri: https://orcid.org/0000-0002-5719-4024
;~assemblyAuthor:
;~  name: David Molik
;~  uri: https://orcid.org/0000-0003-3192-6538
;~dateCreated: '2022-03-21'
;~checksum: md5:7582b26fcb0a9775b87c38f836e97c42
>Contig 1
AAAATCGATCGGCATA...
.
.
.
```

fields provide additional information on how the genome was assembled and distributed.

The `identifier`, `relatedLink`, and `scholarlyArticle` provide the user with links to external information about the genome that is not present in the FASTA file itself. The `identifier` field is used to associate compact URIs (“CURIEs”) with the reference genome. `identifier` is the main field from which a genome assembly can be mapped to a schema or database of an organisation. Both `relatedLink` and `scholarlyArticle` are conventional URLs. `relatedLink` is intended to associate the reference genome with its host site and mirrors, while `scholarlyArticle` is intended to link the genome to its marker paper. All these fields can take multiple values. FHR requires both URLs and identifiers to be publicly accessible and persistent. In particular, taxonomic information (i.e., `taxon`) such as species and isolate / cultivar must be identified using taxonomy URIs registered with identifiers.org in the `taxon uri` as well as the name of the taxon in the `taxon name`. When using CURIEs, FHR recommends that they be registered with identifiers.org so that the user can conveniently find the resource associated with the CURIE. As an alternative to identifiers.org, FHR allows Bioregistry (32) CURIEs to be used. If there is no suitable URI for use in an identifier field, the FHR specifications call for the use of the fields `relatedLink` and `documentation` described later.

The `assemblySoftware`, `instrument`, `accessionID`, and `voucherSpecimen` fields provide additional information on how the genome was generated. `instrument` is a multivalue field that refers to sequencing machines and DNA prep instrumentation used to generate the reference genome. `assembly software` is used to store the name and version of the assembly software used in the creation of the genome assembly, `accessionID` refers to the ID of the genome assembly, and `voucherSpecimen` is used to describe the location of the sequenced material.

Another optional field, `genomeSynonym`, can be used to add one or more common names to the reference genome to supplement the primary `genome` field. It is particularly useful when `genome` contains an unfriendly machine-readable identifier.

`documentation` can be used to provide additional information about the genome. It is free human-readable text that can be used to describe how the reference genome was made, its provenance, usage caveats, or any other information that users of this data should be aware of.

The final optional fields are `funding` and `reuseConditions`. `funding` allows the project funders to be acknowledged and can be multi-valued. The `reuseConditions` field can be used to specify the licencing terms under which genomic data can be used, if any.

Although these fields are all optional, providing them helps increase the FAIRness and TRUSTworthiness of the reference genome. They were chosen to give the authors of the genome multiple complementary ways to document the contents of the file. For example, the `genome` ID field would typically be used to obtain information about the genome from external sources. However, in situations in which the genome’s ID is not sufficient on its own, the authors can add specific relevant information using the `assembly identifier`, `scholarlyArticle`, and `relatedLinks` fields.

Full specification

The full formal specification of FHR is located in our GitHub organisation FAIR-BioHeaders in the repo FAIR-BioHeaders/FHR-Specification. It is versioned using GIT tags. Each release has a full change log, readily accessible through GitHub. Full documentation and examples accompany the FHR-Specification. FHR is specified using a JSON Schema file located at FAIR-BioHeaders/FHR-Specification/fhr.json. The FHR header’s `schema` field points to this file and is used by FHR software tools to verify the formatting and validity of the FHR header. In addition to the JSON schema, FHR has a human-readable specification and examples for FHR. The human-readable specification is located at FAIR-BioHeaders/FHR-Specification and includes both documentation and practical examples.

FASTA header

The recommended method for attaching FHR metadata to a FASTA file is to incorporate it into the FASTA file’s header. The FASTA format consists of a series of one or more nucleotide or protein sequence records, each preceded by a one-line header prefixed with ’>’. The current schema for FASTA provided by NCBI defines the header line specification as consisting of a structured string that includes the Sequence Identifier (SeqID) followed by optional components such as the Resource ID, the accession number and the name of the Sequence (33). There is no formalized way to add additional information to the sequence-level header line (i.e. various resources utilize the sequence header differently).

FHR exploits a legacy feature of FASTA, the FASTA comment, which was included in the original specification created by Lipman & Pearson (1985). The comment is an optional multiline header that appears at the top of the FASTA file, each line of which is preceded by a semicolon (see Table 1). The FHR FASTA header consists of the metadata formatted according to the YAML specification, described below, and each line is prefixed by ”;∼”. This technique provides a format that is both easy to read and easily parsed computationally (34). Furthermore, this format fits into recommendations presented in Batista’s paper on ”Machine actionable metadata models” (35).

The original FASTP tool, which introduced the file format of the same name and was published in 1985, supported FASTA comments (4, 36), and legacy tools that load and parse FASTA files are supposed to ignore these lines. Unfortunately, this is not always the case. Although some modern FASTA-consuming tools recognise and ignore semicolon-based FASTA comments, most do not. Fortunately, it is trivially easy to strip comments out of a FASTA file by removing lines that begin with semicolons. Users of FHR-enabled FASTA files may need to add this preprocessing step to their nucleic acid analysis pipelines before passing the file to downstream tools.

Table 2. FHR Optional Fields

Field	Example	Description
assemblySoftware	Assembler.py	assembly software used in the creation of the genome
documentation	assembly of Example species	documentation about the genome
funding	Project number 1024512128	grant line item, multiple funding lines are allowed
genomeSynonym	Other common names	Other names of the genome
identifier	eg:1024512128	miscellaneous identifiers used to identify the genome
instrument	Awesome Sequencer IIe	physical tools and instruments used
reuseConditions	Creative Commons 4.0	Specifying how the information in the file can be used
accessionID		physical ID of the genome assembly
accessionID name	BioSample:00000000	(property of accessionID, the ID of the sample)
accessionID uri	https://example.org/awesome_science/project-1024	(property of accessionID, the URL of the physical sample)
voucherSpecimen	Located in Freezer 33, Drawer 137 at Cool Science Organisation Rm 1024	a description of the physical sample
relatedLink	https://example.org/example_species/our_genome	related URLs to the genome, multiple relatedLinks are allowed
scholarlyArticle	https://doi.org/10.3390/insects12070626	the genome of the scholarly article was published in

Figure 2. Expanded with optional fields FHR FASTA header example with sequence

```
;~schema: https://raw.githubusercontent.com/FAIR-bioHeaders/FHR-Specification/main/
   fhr.json
;~schemaVersion: 1
;~genome: Example species
;~genomeSynonym: Eg. species
;~taxon:
;~  name: Example species
;~  uri: https://identifiers.org/taxonomy:0000
;~version: 0.0.1
;~metadataAuthor:
;~  name: Adam Wright
;~  uri: https://orcid.org/0000-0002-5719-4024
;~assembler:
;~  name: David Molik
;~  uri: https://orcid.org/0000-0003-3192-6538
;~dateCreated: '2022-03-21'
;~accessionID:
;~  name: Sample:000000
;~  url: https://example.org/awesome_science/sample-1024
;~instrument:
;~- Amazing Sequencer IIe
;~- Neato Sequencer
;~voucherSpecimen: Located in Freezer 33, Drawer 137
;~scholarlyArticle: https://doi.org/10.1371/journal.pntd.0008755
;~documentation: 'Built assembly from... '
;~identifier:
;~- eg:10245122561286432l6842
;~relatedLink:
;~- https://example.org/example-species/our_genome
;~funding: 'some'
;~reuseConditions: 'public domain'
;~checksum: md5:7582b26fcb0a9775b87c38f836e97c42
>Contig 1
AAAATCGATCGGCATA
.
.
.
```

Other serialisation methods

Although the FASTA header format is the preferred implementation of FHR, presenting and transferring FHR header may necessitate other FHR format serialisations. When not embedded directly in the FASTA header, FHR data can be associated with a reference genome by pairing it with a supplementary file that shares the same folder and base name as the FASTA file. This avoids issues with FASTA-consuming software tools that cannot parse the semicolon-delimited comments used by the FASTA FHR header. FHR supplementary files can be represented using several different file formats, JSON, YAML, and HTML, each specialised for a different use case. JavaScript Object Notation (JSON) is widely supported across multiple programming languages, and YAML Ain't Markup Language (YAML) provides a format that is both machine- and human-readable, while Microdata embedded HTML is used to expose metadata to web search engines. We provide a conversion tool to inter-convert these formats.

We recommend naming the FHR file using the template `<genome file name>.fhr.<yaml|json|html>`. It should be located in the same filesystem directory as the FASTA file. In the case of FASTA files that are referenced using a URL, the FHR file should share the same URL path as its corresponding FASTA file. Following a standard naming system reduces the chances that the FASTA metadata becomes decoupled from the data or associated with the wrong data.

Checksum generation method

The checksum is computed based on the contents of the file, minus the checksum, which is injected into the header after being generated. In this way, the checksum can uniquely identify the reference genome sequences, as well as the header in the downstream analysis. For example, for a BAM file generated from an FHR-compliant reference genome, the reference genome can be uniquely identified through its FHR header's checksum, facilitating accurate provenance tracking throughout the analysis pipeline. As all serialisations of FHR are formatted text files, a UNIX command-line program can be used to generate the file's checksum (in Python the *hashlib.md5* command is used). For the validation of the FHR FASTA file with the checksum, FHR provides a command-line tool. An example of the command line tool is in Table 3.

Software support

The schema code has been written to help developers work with FHR-specified reference genomes; the JSON validation tool is used to ensure that the FHR metadata conforms to the specification, a conversion tool to convert the FHR metadata into several different file formats, and the FHR software library used within applications written in Python.

Conversion and validation validation tools

The FHR FASTA header can be validated using the JSON Schema (see: json-schema.org) using the FHR conversion and validation toolkit. This toolkit, which includes the FHR conversion and validation library, is written in Python. The Python library, in turn, depends on the Python libraries JSON, JSON-Schema, microdata, RE, and YAML. FHR tools provide several entry points into the Python library using the command line tools *fhr-convert* and *fhr-validate* and others (see Table 3). *fhr-convert* is for conversion between supported file formats, while *fhr-validate* is for validation of FHR metadata. The formats of the input and output files of the convert tool are automatically determined by the extension of the file. If the header data is stored in a separate file and the user wishes to combine or merge it with its corresponding FASTA file to create a FASTA file with an embedded FHR header, the user can use *fhr-combine-fasta*. To validate the FHR header of a file, use *fhr-validate*. This will check for the presence and syntactic correctness of the header but will not verify that the checksum and the file contents match. To validate the checksum also in an FHR FASTA file, the user would use *fhr-fasta-validate*. In the case of a header/data mismatch, this tool will issue an error message. If the user would like to remove the FHR header from a FASTA file, the *fhr-fasta strip* can be used. Examples of command line tools can be found in Table 3.

DISCUSSION

FHR is designed to facilitate the transfer of provenance and metadata from external sources to or near genome assembly in a way that is minimally disruptive to existing bioinformatics workflows. To meet its goals of enabling the FAIR principles and reducing the risk of genome data and metadata divergence, FHR uses FASTA comments, keeps the included metadata to a minimum, provides a Python processing library, and includes alternative data serialisation. Additionally, FHR is written in Python, relies on JSON-schema for schema validation, and has a small codebase with few dependencies, making it less likely to enter an end-of-life stage in the near or medium future and reducing the maintenance burden. The only core dependency is the JSON schema, which itself relies on a JSON file that describes the schema rules, which is relatively small and can be easily converted if necessary.

Achieving The Design Goals of the FHR format

FHR was designed with ease of implementation and backward compatibility in mind. Despite being described as "Minimal Information" models, many other metadata standards require a large number of fields and relationships, which makes implementation a challenge (see Taylor, 2007 (37) and Brazma, 2001 (38) for examples). Although these comprehensive standards theoretically offer robust inference capabilities, their level of adoption tends to be low due to the complexity of implementation, as information science theory might suggest (39). Following a design philosophy originating in the logic systems field (40, 41), FHR provides a fully featured metadata standard using as few fields as possible to achieve its goals. In an informal assessment of the time and effort needed to generate an FHR (Fragment Header) from the beginning, it was observed that an individual lacking familiarity with an assembly could proficiently produce a comprehensive FHR header within approximately 15 minutes.

Beyond ease of implementation, the FHR specification has four main design objectives:

Table 3. FHR Command Line

<table>
<tr><th>Command</th><th>Description</th></tr>
<tr><td>$ fhr-convert \
<input>.<fasta|json|yaml|html&gt; \
<output>.<fasta|json|yaml|html&gt;</td><td>Converts the input to the output file format which is automatically determined by the extension of the input file.</td></tr>
<tr><td>$ fhr-validate \
<input>.<fasta|json|yaml|html&gt;</td><td>validates the FHR header of any type of input file. This will check for the presence and syntactic correctness of the header, but will not verify that the checksum and the file contents match.</td></tr>
<tr><td>$ fhr-fasta-combine \
<input>.<fasta|json|yaml|html&gt; \
<input>.fasta</td><td>If the header data is stored in a separate file and the user wishes to combine or merge it with its corresponding FASTA file to create a FASTA file with an embedded FHR header</td></tr>
<tr><td>$ fhr-fasta-validate \
<input>.fhr.fasta</td><td>Validates header and checksum. In the event of a header/data mismatch, this tool will issue an error message.</td></tr>
<tr><td>$ fhr-fasta-strip \
<input>.fhr.fasta
<output&gt;</td><td>Removes the FHR header from a FASTA file and writes the result to an output file.</td></tr>
</table>

1. Provide the user with the necessary metadata needed to replicate an analysis performed with a reference genome;

2. Support a variety of text representations;

3. Keep the metadata close to the data; and

4. Enable FAIR and TRUST. Furthermore, to ensure that data standards are beneficial to scientists and other stakeholders, it is necessary to design them to be user-friendly; and compatible with other standards and tools (42). Furthermore, ease of use and compatibility should be proportional to the utility of the new tool(43, 44, 45). Therefore, FHR is designed to fit into existing data and tool ecosystems to be adopted, and because it is located along the reference genome itself, it works more easily in a decentralised data environment.

Provide the metadata necessary to unambiguously identify the provenance of a genome
The FHR specification has been designed to provide the user with the data they need to trace the origin of a genome. This is done in two ways: by using a checksum to ensure the genome remains unchanged as it moves from one platform or device to another and by providing provenance metadata which can be used to reconstruct the continuity of the reference genome. For instance, the `genomeSynonym` field in FHR can be used to identify other names for the genome, or to link to related scholarly articles, authors, and instruments. Additionally, FHR provides information in the header to inform the user of the file’s contents, as well as links to external resources such as the NCBI.

Keep the metadata and data close
FHR ensures that metadata are easily accessible and correctly linked to the data by keeping data and metadata in the same file. The checksum in the FHR header is calculated on the basis of both the data and metadata. This allows the user to accurately identify the metadata and the data used in the analysis when using the checksum to identify the reference genome. This strong connection guarantees the origin of both the metadata and the data used in the analysis.

Support a variety of implementations
The principles of FHR design recognise that there are times when it is preferable to maintain the original FASTA file as is. Therefore, it is essential to offer a range of alternative serialisation techniques so that metadata can be stored alongside the data in the FASTA file. This is also beneficial in certain specialised circumstances, such as search engine optimisation, when we may want the metadata to be visible to a search engine without making the sequence data visible.

Enable FAIR and TRUST
FHR has been designed to ensure that those following the specification meet the core FAIR and TRUST principles. We designed the format with two user groups in mind. The first is the repository staff that provides the reference genomes to the community. The second group consists of researchers who use the reference genomes within their analytical pipelines. By supporting the FHR specification, the data repositories make the metadata more accessible to those using the files and will benefit from the users being able to find the data in the originating repository more readily. At the same time, the analytic pipelines written by researchers will become FAIRer and more TRUSTworthy by enabling the unambiguous identification of the genome(s) in use.

Adoption Challenges

There are many software packages for manipulating FASTA files, but none currently understand FHR headers. Of more concern is the fact that many FASTA libraries do not follow the original standard’s use of the semicolon character to mark comment lines. To address this, the FHR toolset provides a command to strip the FASTA file of the FHR header when using software that cannot handle the header (see Table 3). In the future, we will work with the bioinformatics community to adopt standard pipelines to handle FHR-containing FASTA files. This will involve adding logic to existing FASTA software libraries to handle comments.

An unresolved issue is who has the authority to generate the FHR headers, which is a matter that needs to be debated by the stakeholder research communities. To minimise areas of authorship controversy, in the FHR header, the authorship of the assembly is separated from the authorship of the metadata, allowing a trusted third-party group, such as the curators of a model organism resource, to provide the metadata header.

Time and effort needed to create an FHR header

The time and effort required to develop an FHR header for a typical assembly is an important metric in the ease of adoption. To explore this question, we ran an example workflow in which we selected an unfamiliar mammalian genome assembly from an NCBI reference genome page (mouse GRCm39, RefSeq accession `GCF_000001635.27`), downloaded it, and created an appropriate FHR header using the data available on its NCBI genome page. In testing, the process of writing a single FHR header took 15 minutes, developing a template header JSON file to hold the information provided by NCBI. With automation, we estimate that the time per imported genome would be roughly five minutes.

Related Efforts

FHR is not the only project that attempts to solve the problem of aligning the storage of assembly metadata and the exchange of said data between resources. For example, the Minimum Information about a Genome Sequence (MIGS) specification, published in 2008, provides a set of fields for various types of assemblies with the intent of generating reports that are used to exchange information between resources (46). Compared to FHR, MIGS specification provides minimal information that should be tracked for a reference genome, whereas FHR only provides the fields that should be a header. Future versions of FHR may provide an extension mechanism that allows for the addition of fields from the full MIGS checklist.

FairGenomes is another project that tries to solve problems with metadata in Genomes (47). FairGenomes is designed for personal human genomes used in medical studies and takes advantage of a more extensive schema designed around using the stored metadata for personal human genomes in downstream analysis. FHR and FairGenomes have different design goals and use cases.

Fostering community adoption

Data standards facilitate the rapid spread of concepts and ideas and become more useful the more researchers utilise them (ie, the network effect). To achieve widespread adoption, FHR needs early adopters to start generating and providing FHR-compliant reference genomes, while, in parallel, working on increasing the number of tools that can take advantage of additional information. Although it is still in its early days, the FHR header format has already been picked up and used by several significant biological curation efforts.

USDA ARS Pecan Germplasm phenotype database

The United States Department of Agriculture, Agricultural Research Service, Crop Germplasm Research Unit runs a Pecan Breeding program for pecan farmers, which contains phenotype measurements of living pecan trees and a library of nuts in its collection. The pecan phenotype database is being extended with genotype information that includes reference genomes of the pecan trees, and the metadata for the reference genomes will be determined by and exportable into FHR. In addition to the soft rollout of the Pecan phenotype and genotype database, other USDA projects are also starting to adopt FHR.

Alliance of Genome Resources

The Alliance of Genome Resources(29), a consortium of six Model Organisms and the Gene Ontology Consortium, is currently creating FHR YAML files. FHR YAML files are displayed within their JBrowse (48) genome browser instances.

AgBioData

AgBioData is a collection of Biological Resources with a mission to consolidate standards and is funded by the United States National Science Foundation. They are tasked with generating recommendations by acquiring, displaying, and reusing genomic, genetic, and breeding (GGB) data (49). AgBioData FAIR Scientific Literature and Genome Assembly and Annotation Nomenclature working groups are specifically orientated toward the problems of genome data and the metadata divergence problem.

We are working with this consortium as part of the FAIR Literature Working Group and the Genome Assembly Working Group to adopt FHR headers to use with the large number of genomes generated in agriculture-related projects.

MicroPublications

microPublication Biology is a peer-reviewed journal that revolutionizes the scholarly communication workflow by integrating data validation and curation into the publishing process. This curatorial-driven approach allows editors to review and authenticate domain-specific naming conventions and experimental reporting standards before publication. By working with the microPublication editorial team during publication, we will guide authors in properly reporting genome metadata standards by adopting FHR headers, alleviating numerous obstacles to making these data reusable.

KBase

KBase The Department of Energy Systems Biology Knowledgebase is currently testing its pipelines to see how well they handle files with FHR headers.

ATCC

ATCC's Sequencing & Bioinformatics Center is testing the production of FHR compliant genome assemblies for the ATCC Genome Portal. Implementation into their production pipelines and inclusion in data deliverables to end-users is expected to be completed in 2024.

The goal of FHR is to provide a standard for maintaining the provenance of reference genomes that are translatable between storage and analysis platforms. To this end, FHR provides a schema of the minimal set of fields to facilitate the application of the FAIR and TRUST principles when

Table 4. Software Locations for FHR Relevant Code

GitHub	DOI
FHR-Specification	10.5281/zenodo.6762549
FHR-File-Converter	10.5281/zenodo.6762547
Jbrowse (V1.7.9) with FHR support	

generating and sharing Reference Genome files, as well as conversion and validation tools designed to make the creation, interconversion, and validation of FHR files easier.

CONCLUSION

A metadata standard for reference genomes does not currently exist, and FHR represents a lightweight solution to this problem. FHR provides reference genomes with unambiguous identifiers, a method for validating the contents of the genome, and provenance identifying information, without placing an undue burden on the authors and maintainers of the genome. By providing several alternative implementations, FHR is minimally disruptive while providing many long-term benefits to the bioinformatics community. However, a transition to FHR will require the commitment of time and effort of various genomic data repositories and tool developers before its benefits will be realised by the bioinformatics community.

DATA AVAILABILITY

The code published for FHR is in the public domain per the United States 17 U.S.C. § 105. The code and specification are freely available for use and modification (Table 4).

ACKNOWLEDGEMENTS

The authors thank Natalie Meyers with The Lucy Family Institute for Data and Society at the University of Notre Dame for conversations around research communities focused on metadata standards that were used in the writing of the manuscript. The authors thank Monica Poelchau with the National Agriculture Library and Sarah Dyer at EMBL EBI for relevant discussions.

FUNDING

Adam Wright is supported by the Adaptive Oncology Programme at the Ontario Institute for Cancer Research.

During a portion of this project, David Molik is supported by the USDA Agricultural Research Service (ARS) HQ Research Associate program in Big Data.

A portion of this work was carried out by the Tropical Pest Genetics and Molecular Biology Research Unit, ARS Project number 2040-22430-028-000D.

A portion of this work was carried out by the Arthropod-borne Animal Diseases Research Unit, ARS Project numbers 3020-32000-018-000-D, 3020-32000-020-000-D, and 3020-32000-019-000-D.

This research used resources provided by the SCINet project of the USDA Agricultural Research Service, ARS project number 0500-00093-001-00-D.

We gratefully acknowledge the support of the WormBase grant (U24HG002223), which provided funding for this research. This grant has been instrumental in supporting the contributions of Karen Yook, Daniela Raciti, Paul Sternberg, Adam Wright, Lincoln Stein, and Scott Cain. Their efforts have significantly contributed to the success of this project.

CONFLICT OF INTEREST STATEMENT.

None declared.

REFERENCES

1. Molik, D. C. (07, 2022) An Outsider’s Perspective on Why We Climb Mountains and Why Projects Like the i5k Matter. *Journal of Insect Science,* **22**(4) 2.
2. Schoof, H. (2003) Towards Interoperability in Genome Databases: The MAtDB (MIPS *Arabidopsis Thaliana* Database) Experience. *Comparative and Functional Genomics,* **4**(2), 255–258.
3. Niu, Y. N., Roberts, E. G., Denisko, D., and Hoffman, M. M. (May, 2022) Assessing and assuring interoperability of a genomics file format. *Bioinformatics,*.
4. Lipman, D. J. and Pearson, W. R. (1985) Rapid and Sensitive Protein Similarity Searches. *Science,* **227**(4693), 1435–1441.
5. Jordan, E. and Carrico, C. (1982) DNA database. *Science,* **218**(4568), 108–108.
6. Thorisson, G. A., Muilu, J., and Brookes, A. J. (January, 2009) Genotype-phenotype databases: challenges and solutions for the post-genomic era. *Nat. Rev. Genet.,* **10**(1), 9–18.
7. Brookes, A. J. and Robinson, P. N. (2015) Human genotype–phenotype databases: aims, challenges and opportunities. *Nature Reviews Genetics,* **16**(12), 702–715.
8. Schatz, M. C. (2015) Biological data sciences in genome research. *Genome Research,* **25**(10), 1417–1422.
9. Sousa, R. B., Cugler, D. C., Malaverri, J. E. G., and Medeiros, C. B. (mar, 2014) A provenance-based approach to manage long term preservation of scientific data. In *2014 IEEE 30th International Conference on Data Engineering Workshops* IEEE.
10. Pettengill, J. B., Beal, J., Balkey, M., Allard, M., Rand, H., and Timme, R. (07, 2021) Interpretative Labor and the Bane of Nonstandardized Metadata in Public Health Surveillance and Food Safety. *Clinical Infectious Diseases,* **73**(8), 1537–1539.
11. Herschel, M., Diestelkämper, R., and Ben Lahmar, H. (2017) A survey on provenance: What for? What form? What from?. *The VLDB Journal,* **26**(6), 881–906.
12. Madden, B., Adams, I., Storer, M. W., Miller, E. L., Long, D. D. E., and Kroeger, T., Provenance Based Rebuild: Using Data Provenance to Improve Reliability. Technical Report UCSC-SSRC-11-04, University of California, Santa Cruz (May, 2011).
13. Haslhofer, B. and Klas, W. (mar, 2010) A Survey of Techniques for Achieving Metadata Interoperability. *ACM Comput. Surv.,* **42**(2).
14. Kitts, P. A., Church, D. M., Thibaud-Nissen, F., Choi, J., Hem, V., Sapojnikov, V., Smith, R. G., Tatusova, T., Xiang, C., Zherikov, A., DiCuccio, M., Murphy, T. D., Pruitt, K. D., and Kimchi, A. (11, 2015) Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Research,* **44**(D1), D73–D80.
15. Hirsch, C. and Schildknecht, S. (2019) In Vitro Research Reproducibility: Keeping Up High Standards. *Frontiers in Pharmacology,* **10**.
16. Benton, B., King, S., Greenfield, S. R., Puthuveetil, N., Reese, A. L., Duncan, J., Marlow, R., Tabron, C., Pierola, A. E., Yarmosh, D. A., Combs, P. F., Riojas, M. A., Bagnoli, J., Jacobs, J. L., and Thrash, J. C. (2021) The ATCC Genome Portal: Microbial Genome Reference Standards with Data Provenance. *Microbiology Resource Announcements,* **10**(47), e00818–21.
17. Yarmosh, D. A., Lopera, J. G., Puthuveetil, N. P., Combs, P. F., Reese, A. L., Tabron, C., Pierola, A. E., Duncan, J., Greenfield, S. R., Marlow, R., King, S., Riojas, M. A., Bagnoli, J., Benton, B., Jacobs, J. L., and

Suen, G. (2022) Comparative Analysis and Data Provenance for 1,113 Bacterial Genome Assemblies. *mSphere,* **7**(3), e00077–22.

18. McCarthy, F. M., Gresham, C. R., Buza, T. J., Chouvarine, P., Pillai, L. R., Kumar, R., Ozkan, S., Wang, H., Manda, P., Arick, T., Bridges, S. M., and Burgess, S. C. (November, 2010) AgBase: supporting functional modeling in agricultural organisms. *Nucleic Acids Research,* **39**(suppl_1), D497–D506.
19. Gramates, L. S., Agapite, J., Attrill, H., Calvi, B. R., Crosby, M. A., dos Santos, G., Goodman, J. L., Goutte-Gattat, D., Jenkins, V. K., Kaufman, T., Larkin, A., Matthews, B. B., Millburn, G., Strelets, V. B., and the FlyBase Consortium (03, 2022) FlyBase: a guided tour of highlighted features. *Genetics,* **220**(4) iyac035.
20. Grant, D., Nelson, R. T., Cannon, S. B., and Shoemaker, R. C. (December, 2009) SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Research,* **38**(suppl_1), D843–D846.
21. Colbourne, J. K., Singan, V. R., and Gilbert, D. G. (2005) wFleaBase: the Daphnia genome database. *BMC Bioinformatics,* **6**(1), 45.
22. Davis, P., Zarowiecki, M., Arnaboldi, V., Becerra, A., Cain, S., Chan, J., Chen, W. J., Cho, J., da Veiga Beltrame, E., Diamantakis, S., Gao, S., Grigoriadis, D., Grove, C. A., Harris, T. W., Kishore, R., Le, T., Lee, R. Y. N., Luypaert, M., Müller, H.-M., Nakamura, C., Nuin, P., Paulini, M., Quinton-Tulloch, M., Raciti, D., Rodgers, F. H., Russell, M., Schindelman, G., Singh, A., Stickland, T., Van Auken, K., Wang, Q., Williams, G., Wright, A. J., Yook, K., Berriman, M., Howe, K. L., Schedl, T., Stein, L., and Sternberg, P. W. (02, 2022) WormBase in 2022—data, processes, and tools for analyzing Caenorhabditis elegans. *Genetics,* **220**(4) iyac003.
23. Amos, B., Aurrecoechea, C., Barba, M., Barreto, A., Basenko, E. Y., Bażant, W., Belnap, R., Blevins, A. S., Böhme, U., Brestelli, J., Brunk, B. P., Caddick, M., Callan, D., Campbell, L., Christensen, M. B., Christophides, G. K., Crouch, K., Davis, K., DeBarry, J., Doherty, R., Duan, Y., Dunn, M., Falke, D., Fisher, S., Flicek, P., Fox, B., Gajria, B., Giraldo-Calderón, G. I., Harb, O. S., Harper, E., Hertz-Fowler, C., Hickman, M. J., Howington, C., Hu, S., Humphrey, J., Iodice, J., Jones, A., Judkins, J., Kelly, S. A., Kissinger, J. C., Kwon, D. K., Lamoureux, K., Lawson, D., Li, W., Lies, K., Lodha, D., Long, J., MacCallum, R. M., Maslen, G., McDowell, M. A., Nabrzyski, J., Roos, D. S., Rund, S. S. C., Schulman, S. W., Shanmugasundram, A., Sitnik, V., Spruill, D., Starns, D., Stoeckert, C. J., Tomko, S. S., Wang, H., Warrenfeltz, S., Wieck, R., Wilkinson, P. A., Xu, L., and Zheng, J. (October, 2021) VEuPathDB: the eukaryotic pathogen, vector and host bioinformatics resource center. *Nucleic Acids Research,* **50**(D1), D898–D911.
24. Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., et al. (2021) Ensembl 2021. *Nucleic acids research,* **49**(D1), D884–D891.
25. Oliver, S. G., Lock, A., Harris, M. A., Nurse, P., and Wood, V. (June, 2016) Model organism databases: essential resources that need the support of both funders and users. *BMC Biology,* **14**(1).
26. Sills, J., Robinson, G. E., Hackett, K. J., Purcell-Miramontes, M., Brown, S. J., Evans, J. D., Goldsmith, M. R., Lawson, D., Okamuro, J., Robertson, H. M., and Schneider, D. J. (2011) Creating a Buzz About Insect Genomes. *Science,* **331**(6023), 1386–1386.
27. Levine, R. (apr, 2011) i5k: The 5,000 Insect Genome Project. *American Entomologist,* **57**(2), 110–113.
28. Poelchau, M., Childers, C., Moore, G., Tsavatapalli, V., Evans, J., Lee, C.-Y., Lin, H., Lin, J.-W., and Hackett, K. (October, 2014) The i5k Workspace@NAL—enabling genomic data access, visualization and curation of arthropod genomes. *Nucleic Acids Research,* **43**(D1), D714–D719.
29. Alliance of Genome Resources Consortium (April, 2022) Harmonizing model organism data in the Alliance of Genome Resources. *Genetics,* **220**(4).
30. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data,* **3**.
31. Lin, D., Crabtree, J., Dillo, I., Downs, R. R., Edmunds, R., Giaretta, D., De Giusti, M., L'Hours, H., Hugo, W., Jenkyns, R., Khodiyar, V., Martone, M. E., Mokrane, M., Navale, V., Petters, J., Sierman, B., Sokolova, D. V., Stockhause, M., and Westbrook, J. (2020) The TRUST Principles for digital repositories. *Scientific Data,* **7**(1), 144–.
32. Hoyt, C. T., Balk, M., Callahan, T. J., Domingo-Fernández, D., Haendel, M. A., Hegde, H. B., Himmelstein, D. S., Karis, K., Kunze, J., Lubiana, T., Matentzoglu, N., McMurry, J., Moxon, S., Mungall, C. J., Rutz, A., Unni, D. R., Willighagen, E., Winston, D., and Gyori, B. M. (Nov, 2022) Unifying the identification of biomedical entities with the Bioregistry. *Scientific Data,* **9**(1), 714.
33. Vakatov, D. (2022) The NCBI C++ toolkit book, National Center for Biotechnology Information (US), .
34. Paulk, M., Curtis, B., Chrissis, M., and Weber, C. (08, 1993) Capability Maturity Model, Version 1.1. *Software, IEEE,* **10**, 18–27.
35. Batista, D., Gonzalez-Beltran, A., Sansone, S.-A., and Rocca-Serra, P. (September, 2022) Machine actionable metadata models. *Scientific Data,* **9**(1), 592.
36. Pearson, W. R. and Lipman, D. J. (1988) Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences,* **85**(8), 2444–2448.
37. Taylor, C. F., Paton, N. W., Lilley, K. S., Binz, P.-A., Julian, R. K., Jones, A. R., Zhu, W., Apweiler, R., Aebersold, R., Deutsch, E. W., Dunn, M. J., Heck, A. J. R., Leitner, A., Macht, M., Mann, M., Martens, L., Neubert, T. A., Patterson, S. D., Ping, P., Seymour, S. L., Souda, P., Tsugita, A., Vandekerckhove, J., Vondriska, T. M., Whitelegge, J. P., Wilkins, M. R., Xenarios, I., Yates, J. R., and Hermjakob, H. (2007) The minimum information about a proteomics experiment (MIAPE). *Nature Biotechnology,* **25**(8), 887–893.
38. Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C. P., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., and Vingron, M. (2001) Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nature Genetics,* **29**(4), 365–371.
39. McGuinness, D. L. and Patel-Schneider, P. F. (1998) Usability issues in Description Logic systems. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence* Citeseer.
40. Mcguinness, D. (2005) Ontologies Come of Age.. In *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential* chapter 7, pp. 171–194 MIT Press.
41. Brachman, R. J., McGuinness, D. L., Patel-Schneider, P. F., and Borgida, A. (1999) "Reducing" classic to practice: Knowledge representation theory meets reality. *Artificial Intelligence,* **114**(1), 203–237.
42. Chervitz, S. A., Deutsch, E. W., Field, D., Parkinson, H., Quackenbush, J., Rocca-Serra, P., Sansone, S.-A., Stoeckert, C. J., Taylor, C. F., Taylor, R., and Ball, C. A. (2011) Data Standards for Omics Data: The Basis of Data Sharing and Reuse. In *Methods in Molecular Biology* pp. 31–69 Humana Press.
43. Mangul, S., Martin, L. S., Eskin, E., and Blekhman, R. (February, 2019) Improving the usability and archival stability of bioinformatics software. *Genome Biology,* **20**(1).
44. Shachak, A., Shuval, K., and Fine, S. (October, 2007) Barriers and enablers to the acceptance of bioinformatics tools: a qualitative study. *Journal of the Medical Library Association : JMLA,* **95**(4), 454–458.
45. Kumar, S. and Dudley, J. (May, 2007) Bioinformatics software for biologists in the genomics era. *Bioinformatics,* **23**(14), 1713–1717.
46. Field, D., Garrity, G., Gray, T., Morrison, N., Selengut, J., Sterk, P., Tatusova, T., Thomson, N., Allen, M. J., Angiuoli, S. V., Ashburner, M., Axelrod, N., Baldauf, S., Ballard, S., Boore, J., Cochrane, G., Cole, J., Dawyndt, P., De Vos, P., dePamphilis, C., Edwards, R., Faruque, N., Feldman, R., Gilbert, J., Gilna, P., Glöckner, F. O., Goldstein, P., Guralnick, R., Haft, D., Hancock, D., Hermjakob, H., Hertz-Fowler, C., Hugenholtz, P., Joint, I., Kagan, L., Kane, M., Kennedy, J., Kowalchuk, G., Kottmann, R., Kolker, E., Kravitz, S., Kyrpides, N., Leebens-Mack, J., Lewis, S. E., Li, K., Lister, A. L., Lord, P., Maltsev, N., Markowitz, V., Martiny, J., Methe, B., Mizrachi, I., Moxon, R., Nelson, K., Parkhill, J., Proctor, L., White, O., Sansone, S.-A., Spiers, A., Stevens, R., Swift, P., Taylor, C., Tateno, Y., Tett, A., Turner, S., Ussery, D., Vaughan, B., Ward, N., Whetzel, T., San Gil, I., Wilson, G., and Wipat, A. (May, 2008) The minimum information about a genome sequence (MIGS) specification. *Nature Biotechnology,* **26**(5), 541–547.
47. van der Velde, K. J., Singh, G., Kaliyaperumal, R., Liao, X., de Ridder, S., Rebers, S., Kerstens, H. H. D., de Andrade, F., van Reeuwijk, J., Gruyter, F. E. D., Hiltemann, S., Ligtvoet, M., Weiss, M. M., van Deutekom, H. W. M., Jansen, A. M. L., Stubbs, A. P., Vissers, L. E. L. M., Laros, J. F. J., van Enckevort, E., Stemkens, D., 't Hoen, P. A. C., Beliën, J. A. M., van Gijn, M. E., and Swertz, M. A. (April, 2022) FAIR Genomes metadata schema promoting Next Generation Sequencing data reuse in Dutch healthcare and research. *Scientific Data,* **9**(1).
48. Buels, R., Yao, E., Diesh, C. M., Hayes, R. D., Munoz-Torres, M.,

Helt, G., Goodstein, D. M., Elsik, C. G., Lewis, S. E., Stein, L., and Holmes, I. H. (Apr, 2016) JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biology,* **17**(1), 66.

49. Harper, L., Campbell, J., Cannon, E. K. S., Jung, S., Poelchau, M., Walls, R., Andorf, C., Arnaud, E., Berardini, T. Z., Birkett, C., Cannon, S., Carson, J., Condon, B., Cooper, L., Dunn, N., Elsik, C. G., Farmer, A., Ficklin, S. P., Grant, D., Grau, E., Herndon, N., Hu, Z.-L., Humann, J., Jaiswal, P., Jonquet, C., Laporte, M.-A., Larmande, P., Lazo, G., McCarthy, F., Menda, N., Mungall, C. J., Munoz-Torres, M. C., Naithani, S., Nelson, R., Nesdill, D., Park, C., Reecy, J., Reiser, L., Sanderson, L.-A., Sen, T. Z., Staton, M., Subramaniam, S., Tello-Ruiz, M. K., Unda, V., Unni, D., Wang, L., Ware, D., Wegrzyn, J., Williams, J., Woodhouse, M., Yu, J., and Main, D. (09, 2018) AgBioData consortium recommendations for sustainable genomics and genetics databases for agriculture. *Database,* **2018** bay088.

AUTHORS NOTES

The U.S. Department of Agriculture is an equal opportunity lender, provider, and employer.

Mention of trade names or commercial products in this report is solely to provide specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture.