PLOS COMPUTATIONAL BIOLOGY

EDITORIAL

Ten simple rules for annotating sequencing experiments

Irene Stevens[1,2]*, Abdul Kadir Mukarram[1], Matthias Hörtenhuber[1], Terrence F. Meehan[4], Johan Rung[2,3], Carsten O. Daub[1,2]

1 Department of Biosciences and Nutrition, Karolinska Institutet, Huddinge, Sweden, **2** Science for Life Laboratory, Karolinska Institutet, Stockholm, Sweden, **3** Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden, **4** European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom

* irene.stevens@ki.se

OPEN ACCESS

Citation: Stevens I, Mukarram AK, Hörtenhuber M, Meehan TF, Rung J, Daub CO (2020) Ten simple rules for annotating sequencing experiments. PLoS Comput Biol 16(10): e1008260. https://doi.org/10.1371/journal.pcbi.1008260

Editor: Scott Markel, Dassault Systemes BIOVIA, UNITED STATES

Published: October 5, 2020

Funding: This work has been funded by the EU Horizon 2020 research and innovation programme under the Marie Sklowdowska-Curie grant (No 643062) received by COD. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

A file of nucleic acid sequences itself is not descriptive. Accompanying information describing data, known as metadata, is important for fueling artificial intelligence and ensuring data longevity as technologies evolve. Poor metadata can significantly lower the value of sequencing experiments by limiting the reproducibility of the study and its reuse in integrative analyses. Furthermore, metadata provides the basis for supervised machine learning algorithms using labeled data and indexing Next Generation Sequencing datasets into public repositories to support database queries and data discovery. Thus, metadata is key for making data Findable, Accessible, Interoperable, and Reusable (FAIR) [1].

Several empirical studies have shown the need for better practices in curating scientific data [2–5]. Community efforts to improve metadata quality include various minimum metadata standards such as Minimum Information about a Next-Generation Sequencing Experiment (MINSEQE) [6] or broader principles such as the FAIR guidelines. However, there is a lack of consensus or compliance for many of these standards.

Here, we distilled a few pragmatic principles, which are summarized in Fig 1, to help data producers collect and store high-quality metadata about sequencing experiments. Ultimately, we hope these will increase the resource value of public sequencing data.

Rule 1: Think beyond your initial study question

Metadata is usually specific to a given study, thus the decision of what metadata to collect should be largely determined during the experimental design phase knowing what variables will be created. Think beyond your immediate biological questions, and record everything that systematically varies in the experiment. As early as sample collection, record sufficient descriptive information that will allow others to reproduce your experiment. After sample collection is finished, it will be more difficult to remember sample details, for example, since key personnel might not be present anymore in the lab. Remember to add sufficient details needed to reproduce your study or to support database queries that will discover your data. An example of something which might be missed is information about DNA or RNA fragmentation, sequencing adapter ligation, and library enrichment steps prior to sequencing. Alnasir and colleagues [7] report only 4% of metadata records in the MINSEQE-compliant Sequence Read Archive (SRA) repository contain information about these protocol steps, causing biases in meta-analyses of SRA records.

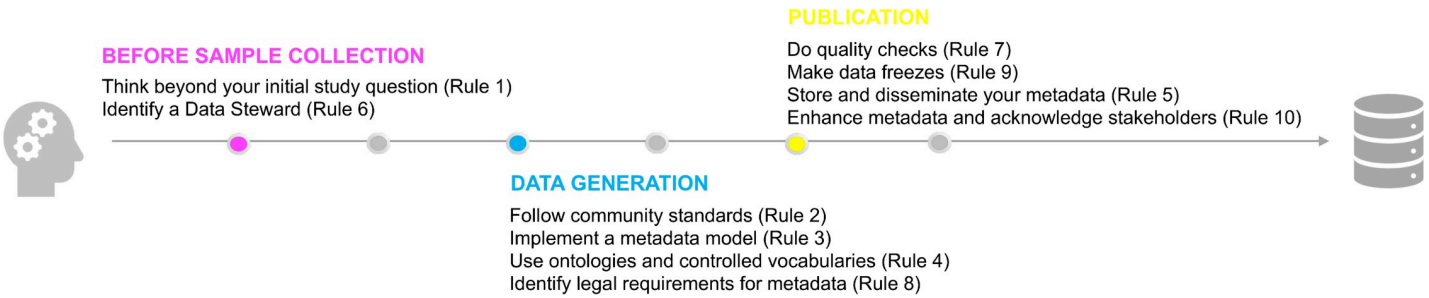

Fig 1. Summary of recommendations for metadata collection at 3 key stages of a sequencing project: before samples collection, during data production, and prior to publication. Note that the rules become increasingly more concrete as the project progresses.

https://doi.org/10.1371/journal.pcbi.1008260.g001

In addition to experimental details, the metadata record should also provide technical details such as barcodes, linkers, and other nucleotide information. Capture the computational aspects such as processing pipelines and the respective software versions. Publish your code and processing environment as a Git repository, Docker container, computational notebook, or Code Ocean capsule. Provide all the code and data needed to reproduce your figures (e.g., count tables). In subsequent rules, we give progressively more concrete ways to design (see Rule 2) and implement (see Rule 3) custom metadata records.

Rule 2: Follow community standards

Meta-analyses, increasingly performed using machine learning approaches, are using metadata to incorporate disparate datasets and find new insights into biological processes. To ensure compatibility of your study with similar studies, adhere to established community standards and formats for metadata and data.

The FAIR guidelines [1] offer high-level advice for making data FAIR. The MINSEQE standard [8] was established by the Functional Genomics Data Society (FGED) similar to the Minimum Information About a Microarray Experiment (MIAME) standard for microarrays [9]. These standards are intended to provide the minimum descriptive information to enable data reuse, and many public repositories are MINSEQE compliant. The Dublin Core Metadata Initiative [10] developed standards and best practice recommendations for creating and sharing metadata, available through the Dublin Core User Guide (dublincore.org/resources/userguide/). The Global Alliance for Genomics and Health (GA4GH) [11] also provides standards and tools for sequencing data, such as the Genomic Data Toolkit (ga4gh.org/genomic-data-toolkit/).

As a first step, determine the minimum standards and requirements of your target repository and journal. Adhering to these requirements is a prerequisite for publishing scientific data. Beyond the minimum standards, it is strongly encouraged to add as much experimental detail as possible.

Rule 3: Implement a metadata model

A metadata model spells out the terms, relationships, and categories used to describe samples and data in a structured manner. One example of a metadata model is the International Human Epigenome Consortium (IHEC) metadata model [12]. Several large-scale sequencing projects, such as the Functional Annotation of the Mammalian Genome (FANTOM5) [13], Encyclopedia of DNA Elements (ENCODE) [14], and the Danio Rerio Encyclopedia of DNA Elements (DANIO-CODE) [15], have established additional metadata models to customarily describe their data in a systematic way that allows for integrative analysis of disparate datasets.

Create a similar metadata specification by listing all the possible terms that will describe your data. Organize terms into progressively broader categories until obtaining only a few umbrella categories that reflect the experimental workflow from sample collection to data processing. Within each category, providing certain terms may be required or optional based on how these are used in downstream analysis.

We previously created a custom metadata specification using a similar approach [16]. We used a top-down structure to capture metadata across the entire experimental workflow from biological sample to library preparation, sequencing procedure, sequencing files, and processed files. We defined 6 metadata sections corresponding to the experiment workflow: Series, Biosample, Assay, Applied Assay, Sequencing, and Data. Under each section, we defined weights on the terms such as required (e.g., biosample type), conditionally required (e.g., target of a chromatin immunoprecipitation sequencing (ChIP-seq assay)), and optional terms (e.g., chemistry version used for sequencing).

The Investigation/Study/Assay Tab-Delimited (ISA-TAB) [17] format is widely used for submitting metadata to repositories. The ISA-TAB format can be implemented as text-based, such as comma-separated values (CSV), tab-separated values (TSV), Excel-based, or relational database depending on the data volume and project resources.

For a smaller sequencing project, it might be useful to take advantage of tools specifically designed for capturing metadata, such as the Center for Expanded Data Annotation and Retrieval (CEDAR) Workbench [18] or ISA-TAB tools [19] (isa-tools.org/index.html). For larger projects, custom implementations can be considered such as the ENCODE Data Coordination Center (DCC) [14] or FANTOM5 Semantic catalogue of Samples, Transcription Initiation, And Regulations (SSTAR) [13].

To help mitigate potential reproducibility issues, consider using workflow management tools (e.g., nf-core [20], Cromwell [21], and Galaxy [22]) and workflow description standards (Common Workflow Language (CWL) [23] and Workflow Description Language (WDL) [21]).

Rule 4: Use ontologies and controlled vocabularies

Maximize the use of ontologies and controlled vocabularies within the metadata fields (see Rule 3). This will reduce misannotations and ensure metadata consistency and compatibility with other datasets. We recommend using a minimum set of ontologies to describe samples (i.e., cell lines, primary cells, and primary tissues), sequencing details (assay types and platforms), or diseases. Useful resources are the Open Biological and Biomedical Ontology (OBO) Foundry [24], National Center for Biomedical Ontology (NCBO) BioPortal [25], or European Bioinformatics Institute (EBI) Ontology Lookup service [26].

When an ontology is not available, consider using controlled vocabulary terms to minimize misannotations in the metadata. For example, create a list of controlled terms such as for file formats (e.g., FASTQ and BAM), for sequencing instruments (e.g., HiSeq X, etc.), or for platforms (Illumina, Ion Torrent, PacBio, etc.) in order to restrict entries to a predefined vocabulary. This will limit the introduction of errors in the metadata record and ease the data input as well.

Rule 5: Store and disseminate your metadata

It is best practice to create a data management plan (DMP) before generating research data [27]. One component of any DMP is the infrastructure for delivery, analysis, and long-term storage of sequencing data and its description. Give careful consideration to the security, data loss prevention, and ease of accessibility for collaborators and analysts. Any metadata that

contains potentially sensitive information should be encrypted and stored in a secure location. Data loss prevention includes measures such as automated backups, storage in multiple locations, and long-term archiving considerations. Metadata should still be easy to share with the research community and collaborators.

Several publicly funded resources are available for long-term archiving and dissemination of sequencing data and accompanying metadata. The National Center for Biotechnology Information database of Genotypes and Phenotypes (NCBI dbGAP) [28] and the European Genome-phenome Archive (EGA) [29] resources specialize in permanent archiving and sharing of personally identifiable genetic and phenotypic data resulting from biomedical research projects including sequencing data. For data that are not personally identifiable, the NCBI SRA [30], the European Nucleotide Archive (ENA) [31], and the DNA Databank of Japan [32] make biological sequence data available to the research community. GEO [33] and BioSamples [34] collect mainly metadata and references to the respective sequencing data in other databases. In addition, institutional repositories (IRs) funded by the host institution may provide additional storage and data dissemination mechanisms as a complement to specialized public sequence repositories. Some examples of IRs are the Science for Life Lab Data Centre (www.scilifelab.se/data/) and the Beijing Institute of Genomics (BIG) Data Center [35].

Consider data and metadata submission requirements when developing a DMP. In case you propose a large-scale project, consider reaching out for input to streamline future submissions.

Rule 6: Identify a data steward

The data production process spans several stages. Thus, metadata collected over an extended time span might not always be complete or consistent. Sometimes, key personnel move on, causing projects to fail moving forward. The best practice is to assign 1 person from the beginning of the project to be responsible for maintaining and periodically reviewing data records. It can be a data manager, a data officer, or any person with data management competence. Ensure this person will stay engaged throughout the life span of the project. This will allow them to identify issues before key personnel move on to other projects. The data steward can also ensure that policy decisions are applied consistently and timely.

Some institutions provide data support, such as information about data policy, help with making DMPs, or e-infrastructure resources. Take advantage of the data resources provided by your institution and ensure compliance with university policies.

Rule 7: Do quality checks

Quality control of sequencing data is important, but it is beyond the scope of this paper. Here, we focus on metadata quality checks as rapid ways to identify inconsistencies and eliminate errors in the metadata. Perform checks systematically as early as the sample collection phase. Beyond that, validate the accuracy of the metadata against the data. For example, a sample is supposed to be male or female, or a certain gene should be knocked out in the sample. More detailed validations can use data-driven methods, such as clustering samples and identifying outliers. Identify and flag missing values, validate entries against accepted ontology or controlled terms, and validate file formats. Avoid recording 0 for missing values, rather use an appropriate flag (e.g., NA). We recommend designing a file naming scheme and discarding poor quality data early to avoid duplication of records. Be clear about the meaning of terms used in describing your data. For example, clearly distinguish technical and biological replicates. Finally, ask the data generator to verify their metadata. Manual curation remains the gold standard for ensuring high-quality metadata.

Rule 8: Identify legal requirements for metadata

Sequencing experiments in human samples raise special ethical and regulatory concerns. The principal investigator is responsible to be aware of and comply with national or regional legal policies applicable to the location where the data are physically stored. Sensitive metadata likewise must comply with domestic and international standards, including the General Data Protection Regulation (GDPR) and Health Insurance Portability and Accountability Act (HIPAA). Verify the requirements of the funding agency, publishing journal, or university for sensitive data. For medical grade sequencing data, additional standards exist, such as ISO13485:2016 or ISO 27001.

Rule 9: Make data freezes

Data changes with time as files are reprocessed, and metadata is corrected or added. A data freeze is a snapshot of raw and processed sequence files, metadata, and computational workflows at specific time points. Large consortium projects such as FANTOM5 [13] and ENCODE [12] manage ever evolving datasets and metadata by performing periodic data freezes. However, any sequencing project, whether large or small, can benefit from freezing data by creating a resource that will never be changed and can be referenced later on. Each freeze captures the state of data in a system that can be used as a reference point for future analyses.

Match major updates throughout the life span of your project by data freezes. In the best case, a freeze documentation (User's Manual) with the version number and time-stamped changelog is created alongside every freeze. Importantly, no modifications may be done to a data freeze, and any changes have to be realized by additional data freezes.

Rule 10: Enhance metadata and acknowledge stakeholders

Enable people to find your data and quickly get an overview before inspecting the metadata spreadsheets or flat files by giving a graphical abstract, summary statistics on data (dataset size, etc.), or provide a track hub for genome browsers.

Finally, the metadata record is a good place to acknowledge contributors to your data, for example, sequencing centers, data centers, funding agencies, etc. Make sure to use the correct identifiers provided by the funding agencies (project grant numbers) or sequencing centers. This will allow research institutions and funding bodies who are parsing metadata to generate summary metrics about the scientific output and impact of the work. It will also ensure continued backing for your institution's support departments.

Conclusion

As sequencing technologies evolve, investigators generate an increasing amount of genomics data. Each sequencing sample may be described by many aspects (metadata) including experimental details, sequencing protocol, and computational steps. This description is directly linked to the longevity and future reuse of sequencing datasets. Here, we distilled some advice on how to address the challenges of high-quality metadata collection for research groups without dedicated data support.

Acknowledgments

We thank the members and contributors of the DANIO-CODE consortium and Dog Genome Annotation (DoGA) consortium.

References

1. Wilkinson MD, Dumontier M, Jan Aalbersberg I, Appleton G, Axton M, Baak A, et al. Addendum: the FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2019; 6(1):6. https://doi.org/10.1038/s41597-019-0009-6 PMID: 30890711
2. Gonçalves RS, Musen MA. The variable quality of metadata about biological samples used in biomedical experiments. Sci Data. 2019; 6:190021. https://doi.org/10.1038/sdata.2019.21 PMID: 30778255
3. Hu W, Zaveri A, Qiu H, Dumontier M. Cleaning by clustering: methodology for addressing data quality issues in biomedical metadata. BMC Bioinformatics. 2017; 18(1):415. https://doi.org/10.1186/s12859-017-1832-4 PMID: 28923003
4. Berrios DC, Beheshti A, Costes SV. FAIRness and usability for open-access omics data systems. AMIA Annu Symp Proc. 2018; 2018:232–241. PMID: 30815061
5. Roche DG, Kruuk LEB, Lanfear R, Binning SA. Public data archiving in ecology and evolution: how well are we doing? PLoS Biol. 2015; 13(11):e1002295. https://doi.org/10.1371/journal.pbio.1002295 PMID: 26556502
6. Taylor CF, Field D, Sansone S-A, Aerts J, Apweiler R, Ashburner M, et al. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nat Biotechnol. 2008; 26(8):889–896. https://doi.org/10.1038/nbt.1411 PMID: 18688244
7. Alnasir J, Shanahan HP. Investigation into the annotation of protocol sequencing steps in the sequence read archive. Gigascience. 2015; 4:23. https://doi.org/10.1186/s13742-015-0064-7 PMID: 25960871
8. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet. 2001; 29(4):365–371. https://doi.org/10.1038/ng1201-365 PMID: 11726920
9. Brazma A, Robinson A, Cameron G, Ashburner M. One-stop shop for microarray data. Nature. 2000; 403(6771):699–700. https://doi.org/10.1038/35001676 PMID: 10693778
10. Deserno TM, Welter P, Horsch A. Towards a repository for standardized medical image and signal case data annotated with ground truth. J Digit Imaging. 2012; 25(2):213–226. https://doi.org/10.1007/s10278-011-9428-4 PMID: 22075810
11. Vis DJ, Lewin J, Liao RG, Mao M, Andre F, Ward RL, et al. Towards a global cancer knowledge network: dissecting the current international cancer genomic sequencing landscape. Ann Oncol. 2017; 28 (5):1145–1151. https://doi.org/10.1093/annonc/mdx037 PMID: 28453708
12. Bujold D, Morais DAL, Gauthier C, Côté C, Caron M, Kwan T, et al. The International Human Epigenome Consortium Data Portal. Cell Syst. 2016; 3(5):496–499.e2. https://doi.org/10.1016/j.cels.2016.10.019 PMID: 27863956
13. Abugessaisa I, Shimoji H, Sahin S, Kondo A, Harshbarger J, Lizio M, et al. FANTOM5 transcriptome catalog of cellular states based on Semantic MediaWiki. Database (Oxford). 2016; 2016.
14. Hong EL, Sloan CA, Chan ET, Davidson JM, Malladi VS, Strattan JS, et al. Principles of metadata organization at the ENCODE Data Coordination Center. Database (Oxford). 2016; 2016.
15. Tan H, Onichtchouk D, Winata C. DANIO-CODE: toward an Encyclopedia of DNA Elements in Zebrafish. Zebrafish. 2016; 13(1):54–60. https://doi.org/10.1089/zeb.2015.1179 PMID: 26671609
16. Hörtenhuber M, Mukarram AK, Stoiber MH, Brown JB, Daub CO.*-DCC: a platform to collect, annotate, and explore a large variety of sequencing experiments. Gigascience. 2020; 9(3).
17. Sansone SA, Rocca-Serra P, Field D, Maguire E, Taylor C, Hofmann O, et al. Toward interoperable bioscience data. Nat Genet. 2012; 44(2):121–126. https://doi.org/10.1038/ng.1054 PMID: 22281772
18. Gonçalves RS, O'Connor MJ, Martínez-Romero M, Egyedi AL, Willrett D, Graybeal J, et al. The CEDAR Workbench: an ontology-assisted environment for authoring metadata that describe scientific experiments. Semant Web ISWC. 2017; 10588:103–110. https://doi.org/10.1007/978-3-319-68204-4_10 PMID: 32219223
19. Rocca-Serra P, Brandizi M, Maguire E, Sklyar N, Taylor C, Begley K, et al. ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. Bioinformatics. 2010; 26(18):2354–2356. https://doi.org/10.1093/bioinformatics/btq415 PMID: 20679334
20. Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, et al. The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol. 2020; 38(3):276–278. https://doi.org/10.1038/s41587-020-0439-x PMID: 32055031
21. Voss K, Gentry J, Van der Auwera G. Full-stack genomics pipelining with GATK4 + WDL + Cromwell. F1000Res. 2017; 6:1379 (poster).
22. Blankenberg D, Coraor N, Von Kuster G, Taylor J, Nekrutenko A. Integrating diverse databases into an unified analysis framework: a Galaxy approach. Database (Oxford). 2011; 2011:bar011.

23. Kotliar M, Kartashov AV, Barski A. CWL-Airflow: a lightweight pipeline manager supporting Common Workflow Language. Gigascience. 2019; 8(7):giz084. https://doi.org/10.1093/gigascience/giz084 PMID: 31321430

24. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007; 25(11):1251–1255. https://doi.org/10.1038/nbt1346 PMID: 17989687

25. Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, et al. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res. 2011; 39:W541–W545. https://doi.org/10.1093/nar/gkr469 PMID: 21672956

26. Côté R, Reisinger F, Martens L, Barsnes H, Vizcaino JA, Hermjakob H. The Ontology Lookup Service: bigger and better. Nucleic Acids Res. 2010; 38:W155–W160. https://doi.org/10.1093/nar/gkq331 PMID: 20460452

27. Everyone needs a data-management plan. Nature. 2018; 555(7696):286. https://doi.org/10.1038/d41586-018-03065-z PMID: 29542698

28. Tryka KA, Hao L, Sturcke A, Jin Y, Wang ZY, Ziyabari L, et al. NCBI's database of genotypes and phenotypes: dbGaP. Nucleic Acids Res. 2014; 42(Database issue):D975–D979. https://doi.org/10.1093/nar/gkt1211 PMID: 24297256

29. Lappalainen I, Almeida-King J, Kumanduri V, Senf A, Spalding JD, Ur-Rehman S, et al. The European Genome-phenome Archive of human data consented for biomedical research. Nat Genet. 2015; 47(7):692–695. https://doi.org/10.1038/ng.3312 PMID: 26111507

30. Kodama Y, Shumway M, Leinonen R, International Nucleotide Sequence Database Collaboration. The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res. 2012; 40(Database issue):D54–D56. https://doi.org/10.1093/nar/gkr854 PMID: 22009675

31. Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tárraga A, Cheng Y, et al. The European Nucleotide Archive. Nucleic Acids Res. 2011; 39(Database issue):D28–D31. https://doi.org/10.1093/nar/gkq967 PMID: 20972220

32. Ogasawara O, Kodama Y, Mashima J, Kosuge T, Fujisawa T. DDBJ Database updates and computational infrastructure enhancement. Nucleic Acids Res. 2020; 48(D1):D45–D50. https://doi.org/10.1093/nar/gkz982 PMID: 31724722

33. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 2013; 41(Database issue):D991–D995. https://doi.org/10.1093/nar/gks1193 PMID: 23193258

34. Courtot M, Cherubin L, Faulconbridge A, Vaughan D, Green M, Richardson D, et al. BioSamples database: an updated sample metadata hub. Nucleic Acids Res. 2019; 47(D1):D1172–D1178. https://doi.org/10.1093/nar/gky1061 PMID: 30407529

35. BIG Data Center Members. The BIG Data Center: from deposition to integration to translation. Nucleic Acids Res. 2017; 45(D1):D18–D24. https://doi.org/10.1093/nar/gkw1060 PMID: 27899658