DATA CURATION NETWORK

FASTA/FASTQ Data Curation Primer

Authors: Laura Bowman, Dryad & Penn State University (ljbowman10@gmail.com); Shannon Sheridan, Pacific Northwest National Laboratory (shannon.sheridan@pnnl.gov); Briana Wham, Penn State University (bde125@psu.edu)

Mentor: Sarah Wright, Cornell University (sjw256@cornell.edu)

Affiliate Contributors (Peer reviewers): Leslie Delserone and Katie Wampole

Suggested Citation: Bowman, Laura; Sheridan, Shannon; Wham, Briana; Wright, Sarah J. (2023). FASTA/FASTQ Data Curation Primer. Data Curation Network. GitHub Repository.

Format Overview

File extensions (FASTA format, *Wikipedia*, 2023)	
.fasta or .fa	standard/generic; most common; sequence only
.fq or .fastq ; commonly compressed to GNU zip format (.gz)	Contains sequence quality scores in addition to the sequence
.fna	nucleic acid
.ffn	nucleotide of gene regions
.faa	amino acid
.frn	non-coding RNA
.txt	open source format

Suggested Citation: Bowman, Laura; Sheridan, Shannon; Wham, Briana; Wright, Sarah J. (2023). FASTA/FASTQ Data Curation Primer. Data Curation Network. GitHub Repository.

Topic	Description
File Extensions	.fasta and .fastq are most common
MIME Type	text/x-fasta
Structure	text-based file format for nucleotide or amino acid sequences
Versions	Introduced in 1985, many versions (e.g. .fasta, .fastq, .fna, .txt)
Primary fields or areas of use	FASTA and FASTQ files are used by research disciplines engaged in bioinformatics and genomics work
Source and affiliation	The original FASTA/Pearson format was released in 1985 and “is described in the documentation for the FASTA suite of programs” (FASTA format, 2023)
Metadata standards	• Minimal Information about a high throughput SEQuencing Experiment (MINSEQE) • MIxS (Minimum Information about any (x) Sequence) standards
Key questions for curation review	• What repository is most appropriate for the type of data being shared? Are there any funder requirements about which repository should be used? • Is there accompanying documentation, and does it contain sufficient details needed to reproduce the study and/or to support database queries that will discover the data (MINSEQE or MIxS at minimum)? • Does this involve any research on information or physical samples taken from human beings? • Is any other sensitive information included, (e.g., location information for endangered species)?
File-level inspection (Adapted from Brown et al., 2018)	• Do you have the expected number of sequence records or files (do they match the file manifest in the readme, or in the accompanying article)? • Are there outliers in sequence lengths? • Are there duplicates in the sequence identifiers? • If FASTQ, do the length of the sequence line (line 2) and of the quality designation line (line 4) match? • Are there repository-specific requirements for formatting or length? • Are there extra or unsupported characters, e.g., tabs in a definition-line or asterisks or periods in the sequence line?

Tools for curation review	• Text editor (or glogg for large files) • NCBI BLAST, or other sequence editors • R, Python, or other scripting languages
Date Created	2023-05-26
Created by	Laura Bowman Shannon Sheridan Briana Wham Sarah J. Wright - Cornell University DCN Mentor - Sarah Wright - Cornell University
Date updated and summary of changes made	Please see README

Background

FASTA and FASTQ are commonly used text-based file formats for storing and sharing nucleotide (DNA or RNA) sequences and/or amino acid (protein) sequences, and are the main focus of this primer. FASTA and FASTQ are the recognized standard file formats for bioinformatics studies, including next-generation

sequencing (NGS), enabling large-scale exchange of data and information associated with massive sequencing projects (Sielemann et al., 2020). NGS refers to high-throughput technologies for large-scale DNA sequencing such as whole genome sequencing, whole-exome sequencing (WES, WXS), RNA-seq, miRNA-seq, ChIP-seq, and DNA Methylation. NGS experiments generate billions of short sequence reads for each sample which when combined with description and annotations can result in files ranging from a few to hundreds of gigabytes (Zhang, 2016). FASTA and FASTQ files can be opened by many sequence alignment applications or text editors. There are various applications that can convert .fasta files.

Description of Format

FASTA

The FASTA file format is the standard text-based format used for nucleotide or amino acid sequences. Sequence data are represented using single-letter codes, and should typically be standard IUB/IUPAC amino acid and nucleic acid codes (Brouwer, n.d.). While FASTA files most frequently use the file extensions “.fa” or “.fasta”, there are, in fact, many variations, with the file extension denoting the type of sequence the file represents. Examples include:

- .fna (nucleic acid)
- .ffna (nucleotide of gene regions)
- .faa (amino acid)
- .frn (non-coding RNA).

A FASTA file can contain one sequence or multiple sequences. If a FASTA file contains multiple sequences, it can be called a “multi-FASTA” file. FASTA files should have a two line format, which may be repeated for files containing multiple sequences:

- The “comment” line, which starts with a “>” character and is followed by basic information about the sequence. Some repositories will have format requirements (e.g., Genbank), however there is no set format that applies to all FASTA files. Summary information often includes a unique accession number, the name of the gene, the type of sequence, and the source organism.
- The second line is the actual sequence letters.

Example FASTA file:

>CL311878.1 282119_LB1-8C13_T7u LBNL-1 Rabbit BAC Library Oryctolagus cuniculus genomic clone LB1-8C13, genomic survey sequence
GCTTTCTTAAGGTGACACCATAAAGGGGGACTTAAGATGGCCTTATGGGGCTGCTGCTGTGGCGCAGTGG
GTTAAAGCTCTGGCCCGAAGTGCGGGCATCCCATATGGGCGCTGGTTCTAGTCCCGGCTGCTCCTCTTCC
AATCCAGCTCTATGCTATGGCCTGGGAAAACAATAGAAAATGGCCCAAGTTCTTGGGCCCCTGCACCCAT
GTGGGAGACCCAGAGGAAGCTCCTGGCTCCTGGCTTTGGATCAGCGCAGCTCCGGCCATTGCGGCCATCT
GGGGAAGGAACCAGCGGATGGAAGACCCCTTTCTGCCTCTGTCTCTCTCTCTCTCTCTAACTTTTTCAAA
TAAATGTATCTTTTTAAAAAAAAAAAAAAAGATGGCTTAGATACAGCGAGAGTTCCCGGATTCTTTTAAC
TTCCTACATCCCAGAAAGACCAAAAATCCAGATGGTGTAGTGGGAAATGCTCTTAGCCCTTGGATGAGAA
AGCCAAATACTAAGTGTAACGATGACATTGTAATGAGCAAGTTTGTGTATTGTTTTCCATGGTAACCTAC
CTATTGGCCATCACCAGTGACCTGAGCTTCCATGAACAATGTCTGTCAATAACAGGCAGACCTCGGGGTT
CTGTGCAGTTGTCAGCTCCACCTCACACTCAGAGCCACAGTTTTGGGAGAAATGGGCACCTATCAAAAGA
CAAAAGGTGAAACATGATTGTACAGGGGACTGAACTGGTGT

(https://www.ncbi.nlm.nih.gov/nuccore/CL311878.1?report=fasta)

FASTQ

The FASTQ format stores the sequence in addition to its quality score (most use the Phred quality score). The Phred quality score is a single-character representation of the quality of the sequence read, or the probability that the sequence is what it has been identified as (Phred quality score, *Wikipedia*, 2023*)*. Both the sequence and the quality score are encoded with a single ASCII character, and there should be the same number of quality score symbols as sequence letters. These files are also plain text, can be represented with the extension ".fq" or ".fastq", and are often compressed to GNU zip format (.gz) Each FASTQ file follows a four line format (FASTQ format, 2023*)*:

- The first line starts "with a "@" character and is followed by a sequence identifier and an optional description"
- The second line consists of the raw sequence letters.
- "The third line begins with a "+" symbol and is optionally followed by the same sequence identifier (and any description again)." The "+" sign serves as a marker indicating the end of the sequence.
- The fourth line contains the quality values for the sequence, and should match the number of letters in the sequence in the second line.

Example FASTQ file:

@ee15a423-b008-44be-a4b2-5441d11b0b94 runid=fa1d76e661ac2bbb53a002e85e75a30e91827c51 sampleid=1 read=5087 ch=53 start_time=2019-10-18T22:14:05Z
GTTGTACTTCGTTCAATCGGTAGGTGTTTAACCGGATGGTCACGCCTACCGTGACAAAGAGATTGTCGGTGTCT
TTGTGTTTCTGTTGGTGCTGATATTGCATTATGCATGAACGTAATGCCCATTAGTTGTGAATCCACCATGCGCGG
AAGATAGAGCGACAGGCAAGTCACAAAGACACCGACAACTGTC
+
##$&$&/035881()'$0&*('-.=;685()$.%($'%%&#&)+..0,&+&%.-/+,%&()$3:0&@09BF=>CC8(78029F7=<=)+@
+.6CCFFC@-8%2579<B8;88412134,,;:8./,#1#&(%((09;B=??48<=<@79*-:B540,8=B=444:<571-B5=ED2.56;
110.5+,*)%%*

(Raposa. (2020). Sample Fastq Files: 1_control_psbA3_2019_minq7.fastq [Data set]. Zenodo. https://doi.org/10.5281/zenodo.3736457)

Diving Deeper into FASTA/FASTQ formats and requirements

- Repositories may have specific requirements and guidelines for the format of the header line, so be sure to check the requirements for the repository in which the data are being deposited. For example, GenBank requires that:
    - "The SeqID must be unique for each nucleotide sequence and should not contain any spaces."
    - SeqID should be 25 characters or less
    - Should only include "letters, digits, hyphens (-), underscores (_), periods (.), colons (:), asterisks (*), and number signs (#)"
    - "Information about the source organism from which the sequence was obtained follows the SeqID and must be in the format [modifier=text]."
    - The sequence title is the final component of the header. "The title should contain a brief description of the sequence" (National Library of Medicine, 2021).
- There may also be repository-specific requirements for the sequence, which is formatted as a one-letter character string. For example, for sequences being deposited in GenBank, the sequence information should only include standard amino acid (U and * are acceptable) and nucleic acid (A,C,G,T) codes and gaps (a single hyphen or dash) or alignment characters. "For sequences that are not contained within an alignment, "?" or "-" characters" are not allowed, and the "IUPAC approved symbol "N" for ambiguous characters" should be used instead (National Library of Medicine, 2021).

- In the sequence itself, it is customary to use separate lines of 60 or 70 characters in length for readability reasons; some repositories will have their own requirements (e.g., GenBank recommends no longer than 80 characters) (National Library of Medicine, 2021).

Sample data set citations

- Citation examples for some of the popular NCBI resources: https://ftp.ncbi.nlm.nih.gov/pub/education/supportcenter/NCBI_services_citation_examples.txt
- Nucleotide [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – . Accession No. CL311878.1, 282119_LB1-8C13_T7u LBNL-1 Rabbit BAC Library Oryctolagus cuniculus genomic clone LB1-8C13, genomic survey sequence.; [2022 December 13]. Available from: https://www.ncbi.nlm.nih.gov/nuccore/CL311878.1

  N.B.! The order of the information is not as important as having sufficient information to find the dataset used.

Key questions to ask yourself

Inspecting the files:

- Can you open the files with text editors available to you? If the files are large, has the researcher included additional documentation about what tools can be used to open them?
- Are there any blank lines in the file? There should not be.
- Are there any hard returns in the FASTA definition line? There should not be.
- Does the file start with a “>” followed by a unique ID?
- Is there a space after the “>”? There should not be.
- Check for repository-specific requirements for header line and/or sequence characters and designation for ambiguous characters, as well as length and size limits.

Human subjects and/or ethics:

- Does this involve any research on information or physical samples taken from human beings that is either subject to IRB or other regulatory approval, used consent forms, or where the data presents ethical quandaries due to human subjects content? If so, see the Human Participants Data Primer. Note that human genomic data may present unique privacy concerns, and that even with de-identified data it may still be possible to identify an individual in the case of whole genome sequencing (Corpas et al., 2018).
- Is there any other sensitive information in the dataset, for example location data of endangered species?

Repositories to consider:

- What repository is most appropriate for the type of data being shared?
  - NIH searchable list of Repositories for Sharing Scientific Data
  - Re3data registry of research data repositories
- Some common data repositories for:
  - Non-human genomic data: Gene Expression Omnibus (GEO), Sequence Read Archive (SRA), Array Express, Mouse Genome Informatics (MGI), WormBase, the Zebrafish Model Organism

Database (ZFIN), GenBank, European Nucleotide Archive (ENA), or DNA Data Bank of Japan (DDBJ).
    - Human genomic data: dpGaP

Metadata and documentation:

- What are you putting elsewhere and what are you putting in an institutional or other data repository? Provide links between related data, code, and articles whenever possible.
- Does the accompanying documentation contain sufficient details needed to reproduce the study and/or to support database queries that will discover the data? While the FASTA/FASTQ format is fairly straightforward, the documentation needed for reuse is more complex. MINSEQE or MIxS standards provide a minimal level of documentation to enable reuse and discovery; however, to make data reproducible, much more information, e.g., methods and workflows used in the analysis, should be included (Brown et al., 2018). Sequencing data require contextual information about the study, sample, experiment, run, and analysis (ten Hoopen et al., 2017). In particular, the sequence analysis methods used after a sample has been collected should be documented to enable others to determine how the sequences could be reused. This information can include the protocol for the extraction or isolation method, elements of MIxS including material processing, nucleic acid extraction, and nucleic acid amplification, as well as sequencing library preparation protocols and parameters including sequencing machine configuration, since each of these steps may have an effect on data quality (ten Hoopen et al., 2017).
- Does the metadata include contributors to the data, for example, sequencing centers, data centers, funding agencies, etc.? If so, use the correct identifiers provided by the funding agencies (project grant numbers) or sequencing centers.

Key clarifications to get from researcher

- Are there any funder requirements about which repository should be used?
- In the case of FASTQ with poor quality scores - are these sequences "worth" sharing? Will it benefit others to have access and attempt to re-use these data?
- What species were the genetic samples collected from?
- Where and when were the species collected from which the genetic samples were collected?
- What software or instrument-specific information is needed to understand or interpret the data, including software and hardware version numbers?
- Licensing

Applicable metadata standard, core elements and readme requirements

The Genomic Standards Consortium (GSC) is a community-driven organization formed in 2005 to establish a **minimum** set of metadata descriptions for genomic data. These are collectively referred to as MIxS (Minimum Information about any (x) Sequence). In addition to these core elements, there are “extensions” available for various types of genomic data which include more than the minimal elements to make the data more FAIR.

- The core descriptors for all genomic data are:
    - Project name
    - Sample name
    - Taxonomy ID of DNA sample
    - Geographic location (latitude & longitude)
    - Geographic location (country and/or sea, region)

    - Collection date
    - Environment (biome, feature, and material)
    - Sequencing method
- The extensions include checklists for genome, metagenome, marker gene, single amplified genome, metagenome-assembled genome, and Uncultivated Virus Genome sequences. The genome sequence and marker gene sequences are further divided into categories to increase specificity.
- Repositories may have their own requirements and standards as well. For example, see the required descriptors for sequencing experiments in the International Nucleotide Sequence Database Collaboration (INSDC) databases, including instrument platform, instrument model, library source, library strategy, and other fields that provide necessary documentation of sequence data context and provenance.
- When possible, use appropriate standards, controlled vocabularies, and/or taxonomies specific to the field of study (e.g., one of hundreds of different biological ontologies searchable in BioPortal or in The Open Biological and Biomedical Ontology (OBO) Foundry (Brown et al., 2018)). In the absence of standards, general readme metadata should accompany the FASTA/FASTQ files, with the core descriptors listed above, along with as much experimental data as possible. Stevens et.al. (2020) state that "[M]anual curation remains the gold standard for ensuring high-quality metadata."

Resources for reviewing data

- NIH Submission Portal
  This portal developed by the NIH helps researchers determine which NCBI repository might be the best option based on the type of genomic data being gathered. Before depositing data into an institutional repository, it is good practice to check whether data should be submitted to one of these databases first. In some instances, these repositories will also ask researchers before submission to confirm that the data are a match for the repository and will be accepted.FAIRsharing (https://fairsharing.org/) is another resource that can be used to navigate a wider range of repositories than solely NCBI-related ones.

Software for viewing or analyzing data

- Text editors can be used to view. However, be aware that some sequences may be too large, and a sequence can be corrupted if you view and modify it in the text editor. In these cases, there is a multi-platform GUI application to browse and search through long or complex files called glogg, which is free to download and can be used to load these large sequences.
- BLAST, or Basic Local Alignment Search Tool, is a program that finds regions of similarity between biological sequences. It is widely used in bioinformatics and is one of the main resources used to analyze genetic sequences. In addition to using it to identify similarities among sequences, curators can also use this tool to validate the format of FASTA sequences. FASTA files are one of the accepted inputs to the BLAST tool, which has a web interface and a quick start guide for the new user.
- Manipulation and parsing of FASTA files can be done using standard text-processors and script-based programming languages such as R, Python, Perl, and Ruby (Brouwer, n.d.).

Preservation actions

- FASTA/FASTQ are text-based formats recommended for preservation.
- The MINSEQE established by the Functional Genomics Data Society (FGED) and MIxS standards are intended to provide the minimum descriptive information to enable data reuse. Many public

repositories are MINSEQE compliant, and FASTA/FASTQ files should be accompanied by this minimal level of documentation. Consider adding as much experimental detail as possible to maximize potential for reuse.

What to look for to make sure this file meets FAIR principles

- Findable: quality metadata to enable discovery, including a descriptive title and details about the specimen, experimental methods; appropriate repository for the data; link(s) to publication(s) (DOIs, etc.)
- Accessible: Are the data in an appropriate repository? Using specialist rather than generalist data repositories for genetic data can increase the odds of long-term preservation and access (Corpas et al., 2018), as well as making it more findable.
- Interoperable: FASTA/FASTQ format is more interoperable than other more proprietary formats; encourage conversion to FASTA/FASTQ.
- Reusable: Are the FASTA files accompanied with sufficient documentation to enable reuse (e.g., data quality, study methods, preferably beyond just MINSEQE or MIxS compliant documentation)? FASTA files alone do not contain sufficient documentation for reuse.

Bibliography

- Brouwer, A. C., with contributions by Nathan L. Brouwer (n.d.). *Chapter 16 introducing fasta files | a little book of r for bioinformatics 2. 0*. Retrieved May 25, 2023, from https://brouwern.github.io/lbrb/introducingFASTA.html
- Brown, A. V., Campbell, J. D., Assefa, T., Grant, D., Nelson, R. T., Weeks, N. T., & Cannon, S. B. (2018). Ten quick tips for sharing open genomic data. *PLOS Computational Biology*, *14*(12), e1006472. https://doi.org/10.1371/journal.pcbi.1006472
- Corpas, M., Kovalevskaya, N. V., McMurray, A., & Nielsen, F. G. G. (2018). A FAIR guide for data providers to maximise sharing of human genomic data. *PLOS Computational Biology*, *14*(3), e1005873. https://doi.org/10.1371/journal.pcbi.1005873
- FASTA format. (2023, August 23). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=FASTA_format&oldid=1144411708
- FASTQ format. (2023, August 28). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=FASTQ_format&oldid=1159281216

- National Library of Medicine. (2021, February 24). *Fasta format for nucleotide sequences*. (n.d.). Retrieved July 12, 2023, from https://www.ncbi.nlm.nih.gov/genbank/fastaformat/
- Phred quality score. (2023, July 10). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Phred_quality_score&oldid=1164616985
- Sielemann, K., Hafner, A., & Pucker, B. (2020). The reuse of public datasets in the life sciences: Potential risks and rewards. *PeerJ*, *8*, e9954. https://doi.org/10.7717/peerj.9954
- Stevens, I., Mukarram, A. K., Hörtenhuber, M., Meehan, T. F., Rung, J., & Daub, C. O. (2020). Ten simple rules for annotating sequencing experiments. *PLOS Computational Biology*, *16*(10), e1008260. https://doi.org/10.1371/journal.pcbi.1008260
- Ten Hoopen, P., Finn, R. D., Bongo, L. A., Corre, E., Fosso, B., Meyer, F., Mitchell, A., Pelletier, E., Pesole, G., Santamaria, M., Willassen, N. P., & Cochrane, G. (2017). The metagenomic data life-cycle: Standards and best practices. *GigaScience*, *6*(8). https://doi.org/10.1093/gigascience/gix047
- Zhang, H. (2016). Overview of sequence data formats. *Methods in Molecular Biology (Clifton, N.J.)*, *1418*, 3–17. https://doi.org/10.1007/978-1-4939-3578-9_1

Additional Recommended Reading

- Brouwer, A. C., with contributions by Nathan L. Brouwer (n.d.). *Chapter 16 introducing fasta files | a little book of r for bioinformatics 2. 0*. Retrieved May 25, 2023, from https://brouwern.github.io/lbrb/introducingFASTA.html
- Brown, A. V., Campbell, J. D., Assefa, T., Grant, D., Nelson, R. T., Weeks, N. T., & Cannon, S. B. (2018). Ten quick tips for sharing open genomic data. *PLOS Computational Biology*, *14*(12), e1006472. https://doi.org/10.1371/journal.pcbi.1006472