Rīga Stradiņš University
Faculty of Medicine
Master's study programme "Biostatistics"

MASTER'S THESIS

Creating a FAIR-Compliant Metadata Framework for RNA Sequencing: A Case Study on Amyotrophic Lateral Sclerosis

Author:
Dr. med. Edīte Vārtiņa
Student ID No 23-001589

Supervisor:
Dr. rer. nat. Baiba Vilne,
Head of Bioinformatics Group, RSU

Riga, 2025

Abstract

The reproducibility "crisis" in biomedical research has underscored the critical need for well-annotated high-quality metadata, particularly in high-throughput domains such as RNA sequencing (RNA-seq). RNA-seq has become an essential technique for investigating gene expression, alternative splicing, and transcriptomic dynamics in various biological contexts, including complex diseases. However, despite the increasing availability of large-scale transcriptomic datasets in public repositories such as Gene Expression Omnibus (GEO), ArrayExpress, and the Sequence Read Archive (SRA), incomplete, inconsistent, or poorly standardized metadata significantly limits their interpretability, reproducibility, and reuse. These issues not only compromise data quality but also hinder meaningful cross-study comparisons, data integration, and meta-analyses.

This study addresses these challenges by developing a FAIR-compliant metadata framework tailored to RNA-seq experiments. The framework is grounded in the FAIR principles, Findability, Accessibility, Interoperability, and Reusability, and is designed to support reproducible, transparent, and machine-readable data annotation. To construct the framework, four widely adopted metadata standards were systematically analyzed: MIAME (Minimum Information About a Microarray Experiment), MINSEQE (Minimum Information about a High-Throughput Nucleotide Sequencing Experiment), FAANG (Functional Annotation of Animal Genomes), and HCA (Human Cell Atlas metadata guidelines). MIAME and MINSEQE were originally published as narrative descriptions without formalized templates. For this study, those descriptive guidelines were translated into structured templates through the identification and categorization of metadata fields relevant to RNA-seq workflows. In contrast, FAANG and HCA already provide well-defined templates with rule-based metadata fields and ontology integration. These existing templates were adopted without modification to maintain compatibility and demonstrate the feasibility of harmonizing diverse metadata schemas under a unified framework.

A metadata template was designed to guide researchers through the annotation process using fill-in fields and multiple-choice options, improving usability while supporting automation. The framework was implemented in Microsoft Excel and incorporates features such as drop-down menus, conditional formatting, and controlled vocabulary fields to enhance accuracy and consistency. To illustrate practical applicability, the framework was applied and evaluated through a case study on RNA-seq data from the DC4MND project, which investigates transcriptomic changes in motor neurons of ALS model mice following trans-spinal direct current stimulation (tsDCS). Results demonstrated that the framework effectively captures key

biological and technical metadata, facilitates compliance with FAIR principles, and enables accurate linkage between samples and data files across complex, multi-center studies. The study emphasizes the importance of prospective metadata curation, especially for technically detailed fields, and outlines plans to develop a user-friendly software tool based on this framework to support wider adoption and scalability. By providing a structured and FAIR-aligned approach to RNA-seq metadata annotation, this framework has the potential to significantly improve data quality, enhance transparency in biomedical research, and support the broader goals of open science and reproducibility.

Anotācija

FAIR prasībām atbilstošas metadatu struktūras izveide RNS sekvencēšanai: Amiotrofiskās laterālās sklerozes gadījuma izpēte

Reproducējamības "krīze" biomedicīnas pētniecībā ir īpaši parādījusi nepieciešamību pēc precīziem, konsekventiem un labi strukturētiem metadatiem, sevišķi tādās tehnoloģiju jomās kā RNS sekvenēšana (RNA-seq). RNA-seq ir kļuvusi par vienu no būtiskākajām metodēm gēnu ekspresijas, alternatīvās saplūšanas (*splicing*) un transkriptomu dinamikas izpētei dažādos bioloģiskos kontekstos, tostarp kompleksu slimību gadījumos. Tomēr, neskatoties uz to, ka publiskajās datu krātuvēs pieejamo liela mēroga transkriptomu datu apjoms pieaug, metadatu nepilnīgums, nekonsekvence vai vāja standartizācija būtiski ierobežo šo datu interpretāciju, reproducējamību un atkārtotu izmantošanu. Tas ne tikai apdraud datu kvalitāti, bet arī būtiski apgrūtina savstarpēji salīdzināmu pētījumu veikšanu, datu integrāciju un metaanalīzes.

Šis pētījums risina minētās problēmas, izstrādājot FAIR principiem atbilstošu metadatu struktūru, kas īpaši piemērota RNA-seq eksperimentiem. Izstrādātais ietvars balstās uz FAIR principiem, atrodams (*Findable*), pieejams (*Accessible*), savietojams (*Interoperable*) un atkārtoti izmantojams (*Reusable*), un ir veidots, lai atbalstītu reproducējamus, pārskatāmus un mašīnlasāmus datu anotācijas procesus. Lai izveidotu šo ietvaru, sistemātiski tika analizēti četri plaši pielietoti metadatu standarti: MIAME (*Minimum Information About a Microarray Experiment*), MINSEQE (*Minimum Information about a High-Throughput Nucleotide Sequencing Experiment*), FAANG (*Functional Annotation of Animal Genomes*) un HCA (*Human Cell Atlas*) metadatu vadlīnijas.

MIAME un MINSEQE ir publicētas kā aprakstošas vadlīnijas, nenodrošinot konkrētu metadatu struktūras paraugu. Šī pētījuma ietvaros vadlīniju apraksti tika pārvērsti strukturētās veidnēs, identificējot un sistematizējot metadatu laukus, kas būtiski RNA-seq darba plūsmām. Savukārt FAANG un HCA jau piedāvā skaidri definētas veidnes ar noteiktiem metadatu laukiem un integrētām ontoloģijām. Šīs vadlīniju struktūras tika izmantotas nemainītā formā, lai saglabātu saderību un demonstrētu iespēju harmonizēt dažādus metadatu standartus vienotā ietvarā.

Lai atvieglotu pētniekiem anotācijas procesu, tika izstrādāta strukturēta metadatu veidne, kas ietver ievadlaukus, izvēlnes un kontrolētās vārdnīcas. Tā uzlabo lietojamību un atbalsta metadatu ievades automatizāciju.

Lai demonstrētu izstrādātās shēmas praktisko pielietojamību, tā tika piemērota un novērtēta, izmantojot RNA-seq datus no DC4MND projekta, kurā tiek pētītas transkriptoma izmaiņas motoneironos ALS peles modelī pēc transspinālās tiešās strāvas stimulācijas (tsDCS). Rezultāti parādīja, ka ietvars efektīvi aptver būtiskākos bioloģiskos un tehniskos metadatus, veicina atbilstību FAIR principiem un ļauj precīzi sasaistīt paraugus ar datu failiem sarežģītos, daudzcentru pētījumos.

Nodrošinot strukturētu un FAIR principiem atbilstošu pieeju RNA-seq metadatu anotēšanai, šis metadatu ietvars var būtiski uzlabot datu kvalitāti un veicināt atvērtās zinātnes reproducējamību.

Table of Content

Abbreviations

ALS	Amyotrophic Lateral Sclerosis
CAP	Cell Annotation Platform
CO	The Cell Ontology
DC4MND	Multidimensional mechanistic investigations of trans spinal direct current stimulation in motor neuron disease
DO	The Human Disease Ontology
FAANG	Functional Annotation of Animal Genomes
FAIR	Findable, Accessible, Interoperable, and Reusable
GEO	Gene Expression Omnibus
HCA	Human Cell Atlas
HCA	The Human Cell Atlas
MGED	Microarray Gene Expression Database
MGI	Mouse Genome Informatics
MIAME	Minimum Information About a Microarray Experiment
MINSEQE	Minimum Information about a High-throughput Nucleotide Sequencing Experiment
PATO	The Phenotype And Trait Ontology
RNA-seq	Ribonucleic acid sequencing
SRA	The Sequence Read Archive

Introduction

RNA sequencing (RNA-seq) is a powerful technique used to capture the transcriptome of an organism`s cells and tissues, providing insights into gene expression, alternative splicing, and other regulatory mechanisms that govern cellular processes (Smail and Montgomery, 2024). However, RNA-seq generates highly complex data, including millions of short sequence reads, many of which are aligned to specific genes or genomic regions (Deshpande *et al.*, 2023). Due to its complexity, RNA-seq data requires a robust metadata framework to ensure proper interpretation and reuse (Matsuda, 2021).

To address the challenges of data management and reproducibility, the scientific community has adopted the FAIR principles - Findability, Accessibility, Interoperability, and Reusability (Wilkinson *et al.*, 2016). These guidelines ensure that scientific data can be reliably located, accessed, integrated, and reused by both humans and machines. Among the most critical components supporting FAIR principles are the assignment of persistent identifiers, such as Digital Object Identifiers (DOIs), and the inclusion of comprehensive metadata (Juty *et al.*, 2020). Metadata provides descriptive information that explains the context, origin, structure, and intended use of a dataset (van Gils, 2023). In biomedical research, metadata typically outlines the key characteristics of the data, including how it was collected, under what experimental conditions, and which parameters were applied during its generation (Seep et al., 2024). Metadata describing an RNA-seq experiment should cover all essential aspects of the study, from basic contact information and a clear description of the experimental design to detailed sample information and technical specifications of the sequencing process (Brazma *et al.*, 2012). It should also establish unambiguous links between biological samples and associated data files, and include references to the annotation sources and quantification tools used during analysis (Matsuda, 2021). Without this level of detail, even high-quality RNA-seq data may be difficult to interpret correctly or reuse in future research.

Currently, the process of documenting metadata is often manual, inconsistent, and time-consuming, making it difficult to ensure the reproducibility and reusability of the datasets (Gonçalves and Musen, 2019). Moreover, the lack of systematic metadata validation and absence of standardized ontologies contribute to widespread data quality issues in public repositories, which in turn limits the reliability of downstream analyses (Caliskan, Dangwal and Dandekar, 2023). A large-scale analysis of metadata quality in major repositories such as the National Center for Biotechnology Information BioSample and the European Bioinformatics Institute BioSamples found that most records fail to meet basic quality criteria, including

completeness, use of controlled vocabularies, and structural consistency (Gonçalves and Musen, 2019). The study revealed that up to 97% of attribute names used in BioSample metadata were user-defined and not standardized, and a significant portion of ontology-based fields did not contain valid terms, severely hampering data discoverability and reuse. A large proportion of RNA-seq datasets, particularly those submitted directly to The Sequence Read Archive, lack essential metadata, such as library preparation protocols, biological replicates, or processed data files, making them difficult to interpret or reuse (Rustici *et al.*, 2021)

Given these broad challenges in RNA-seq data analysis, the implications become even more pronounced when examining the field of complex diseases, such as Amyotrophic Lateral Sclerosis (ALS), where data is often scattered across multiple platforms and formats (Baxi *et al.*, 2022). ALS is a devastating neurodegenerative disease with poorly understood mechanisms, and RNA-seq may represent a valuable tool for identifying potential biomarkers and therapeutic targets (Grima *et al.*, 2023). However, ALS datasets are often dispersed across research groups, platforms, and sequencing technologies (Baxi *et al.*, 2022). Many datasets lack standardized metadata, making it difficult to compare results, share data across platforms, or integrate the findings from multiple studies (Sheffield, LeRoy and Khoroshevskyi, 2023). The lack of consistent metadata documentation not only hinders the reproducibility of ALS research but also limits the ability to leverage existing data for new insights or therapeutic discoveries (Ziff *et al.*, 2023). A recent analysis of RNA-seq publications found that essential methodological information, such as data processing steps, software versions, and parameter settings, was often missing or underreported, with only 25% of the studies fully describing their computational workflows (Simoneau *et al.*, 2021). As emphasized by recent studies, missing or mislabeled metadata, such as incorrect sample sex or tissue type, has already led to the propagation of erroneous conclusions in high-profile publications, underscoring the need for more rigorous metadata integrity checks (Toker, Feng and Pavlidis, 2016; Caliskan, Dangwal and Dandekar, 2023). In many cases, RNA-seq datasets are submitted without sufficient contextual information, such as experimental variables or data processing protocols, making it difficult for researchers to assess their relevance or apply them in meta-analyses (Rustici *et al.*, 2021).

To enable robust and reproducible RNA-seq analyses in ALS and other complex diseases, metadata must be curated using controlled vocabularies, systematically validated, and interoperable across research platforms. A strategy proposed to address these issues is the adoption of predefined metadata templates with standardized field names and controlled vocabularie (Caliskan, Dangwal and Dandekar, 2023). This structured approach, combined with automated validation tools, can improve the accuracy and interoperability of metadata and

support compliance with FAIR data principles. Without reliable metadata, data reuse, cross-study comparisons, and foundational goals of open science are compromised.

Aim of the study

The aim of this research is to establish a FAIR-compliant metadata framework for RNA-seq studies, ensuring that metadata is complete, consistent, accessible, and reusable.

Objectives

1. To evaluate existing metadata standards relevant to RNA sequencing
2. To compare and integrate existing metadata guidelines and ontologies into a unified FAIR-compliant metadata framework for RNA-seq studies
3. To apply and validate the framework on RNA-seq data related to amyotrophic lateral sclerosis

1. Review of Literature

1.1 The Reproducibility Crisis in Scientific Research and the FAIR Principles

In recent years, increasing concern about the reproducibility of scientific results has led to international efforts aimed at improving the reliability, transparency, and reusability of biomedical research (Cobey *et al.*, 2024). Multiple large-scale surveys and initiatives have shown that many researchers struggle to replicate findings, even within their own laboratories, raising questions about the credibility and usability of published data (Baker, 2016; Cobey *et al.*, 2024). These concerns are particularly pressing in fields such as genomics and transcriptomics, where complex datasets and analytical pipelines increase the risk of errors or inconsistencies (Wong, 2019). One striking example is the issue of sample misannotation in public transcriptomic datasets. A large-scale analysis of human transcriptomics datasets revealed that nearly 50% of studies contained at least one mislabeled sample, with discrepancies in basic metadata such as biological sex (Toker, Feng and Pavlidis, 2016).

In response to growing challenges related to data reuse, reproducibility, and integration, the FAIR Data Principles were introduced as a guiding framework for scientific data management. FAIR stands for Findable, Accessible, Interoperable, and Reusable, and aims to ensure that data are not only made available but are also structured and described in ways that make them actionable by both humans and machines (Wilkinson *et al.*, 2016). These principles are particularly relevant in high-throughput fields such as RNA sequencing (RNA-seq), where large and complex datasets require consistent, rich metadata for effective interpretation and reuse (Gundersen *et al.*, 2021). The table (Table 1) below summarizes the FAIR principles and their corresponding subprinciples, as formulated by Wilkinson et al. (2016), which serve as a foundation for improving the management and reuse of complex datasets such as RNA-seq.

Table 1

The FAIR Guiding Principles (Wilkinson *et al.*, 2016)

Principle	Subprinciple	Description
Findable	F1	(Meta)data are assigned a globally unique and persistent identifier.
	F2	Data are described with rich metadata.
	F3	Metadata clearly and explicitly include the identifier of the data they describe.
	F4	(Meta)data are registered or indexed in a searchable resource.

Principle	Subprinciple	Description
Accessible	A1	(Meta)data are retrievable by their identifier using a standardized communication protocol.
	A1.1	The protocol is open, free, and universally implementable.
	A1.2	The protocol allows for authentication and authorization, where necessary.
	A2	Metadata are accessible, even when the data are no longer available.
Interoperable	I1	(Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
	I2	(Meta)data use vocabularies that follow FAIR principles.
	I3	(Meta)data include qualified references to other (meta)data.
Reusable	R1	(Meta)data are richly described with a plurality of accurate and relevant attributes.
	R1.1	(Meta)data are released with a clear and accessible data usage license.
	R1.2	(Meta)data are associated with detailed provenance.
	R1.3	(Meta)data meet domain-relevant community standards.

Although the FAIR principles provide a solid theoretical foundation, their practical implementation in genomic databases has proven inconsistent. A study by Feijoó et al. (2020) introduced the Bio FAIR Evaluator Framework, which evaluates the FAIRness of genomic databases from both human and machine perspectives (Feijoó *et al.*, 2020). Their findings revealed that while many databases superficially align with the FAIR principles, significant gaps exist, particularly in the interoperability and reusability. Issues such as the lack of standardized vocabularies, incomplete metadata fields, and the absence of persistent identifiers were common, indicating that many genomic resources do not fully satisfy the specific requirements defined by the FAIR subprinciples. The authors argue that FAIR compliance must be assessed using well-defined metrics and automated tools to provide actionable feedback and drive improvements in metadata quality and accessibility (Feijoó *et al.*, 2020).

1.2 Metadata in Biomedical Research

Metadata is data that describes and provides context for other (primary) data, such as its origin, structure, and usage (van Gils, 2023). In the context of biological and biomedical research, metadata refers to the information that describes the essential properties of primary data, such as the experimental setup, conditions, and parameters under which the data were generated (Seep *et al.*, 2024). This includes, but is not limited to, details about sample preparation, data collection methods, instruments used, and any contextual or experimental

variables that may influence the results. As recent surveys confirm, researchers increasingly recognize the foundational role of metadata in reproducibility and reuse, yet still face challenges ensuring its quality and completeness (Cobey *et al.*, 2024).

Complete and well annotated metadata shared on public repositories ensures the reproducibility and transparency of biomedical research and enables effective secondary analysis (Huang *et al.*, 2022). Metadata is crucial for understanding not only the content of the data but also the methodology behind its generation, which can influence its interpretation and usability in future studies.

While metadata is essential for ensuring the usability and reproducibility of data, a significant issue remains with the completeness of metadata shared in public repositories. A systematic study conducted by Huang et al. (2021) assessed the completeness of public metadata accompanying omics data across 137 studies covering six therapeutic areas, including Alzheimer's disease, cardiovascular diseases, and inflammatory bowel disease. The study found that, although 72.8% of metadata was available in the original publications, only 48.6% was fully documented in public repositories (Huang *et al.*, 2022). The discrepancy underscores a significant challenge in making metadata truly accessible and reusable, particularly for high-throughput datasets like RNA-seq.

Another study conducted by Gonçalves et al. (2017) systematically assessed the quality of metadata in the BioSample repository, which contains metadata related to biological samples used in biomedical research. The study found numerous anomalies in the metadata records, including incomplete fields, the use of non-standardized field names, and inconsistencies in the values entered. Specifically, 15% of metadata fields in BioSample used field names that were not specified in the repository's data dictionary, which undermined the interoperability of the data (Gonçalves *et al.*, 2017). One of the most notable issues was the lack of ontology term use for many metadata fields. While BioSample's metadata documentation specifies the use of controlled vocabularies and ontologies for certain fields (such as disease and phenotype), only a small percentage of metadata fields adhered to these standards. As a result, many fields contained inconsistent, inaccurate, or poorly specified values, which made it difficult to integrate and analyze data across different studies. For example, attributes like disease were often populated with terms that did not match any entries in the relevant biomedical ontologies (e.g., Human Disease Ontology). This lack of standardization can create significant barriers to the interoperability of datasets and limit their reusability for future studies (Gonçalves *et al.*, 2017).

Furthermore, the study revealed that Boolean attributes, which require values like "True" or "False", often contained a wide range of invalid values such as "Yes," "No," or even "N/A." Only 27% of the Boolean attributes were well-specified, suggesting that even basic data fields suffer from inconsistent documentation (Gonçalves *et al.*, 2017). This highlights the importance of enforcing data validation mechanisms and using controlled terms to ensure metadata consistency and accuracy. Even more, survey results from the biomedical research community underline that such inconsistencies are often compounded by institutional gaps in training and infrastructure for metadata curation (Cobey *et al.*, 2024).

Beyond completeness and consistency, metadata integrity has emerged as a foundational requirement for credible and reproducible bioinformatics research (Caliskan, Dangwal and Dandekar, 2023). Metadata integrity ensures that datasets not only exist but are accurately described, error-free, and usable across multiple contexts. When metadata is unreliable, even high-quality data may be rendered unusable. Real-world consequences of poor metadata quality are well documented. For example, Toker et al. (2016) identified cases where human lung RNA-seq samples were reused across different COVID-19 studies but annotated with conflicting metadata including different donor sex or tissue origin. These discrepancies passed through peer review and were published in high-impact journals, highlighting the urgent need for automated and systematic metadata validation processes (Toker, Feng and Pavlidis, 2016).

The systems used to submit metadata vary widely in terms of usability and quality control (Bhandary *et al.*, 2018). Inconsistent submission practices, combined with limited validation and oversight, often result in metadata that is difficult to interpret or align with other datasets. This significantly reduces the scientific value of large-scale omics data collections and presents barriers to data reuse and reproducibility. Rajesh et al. (2021) found that essential phenotypic information, such as age, ethnicity, and disease severity, was frequently missing from public genomic datasets, thereby limiting their utility for secondary research (Rajesh *et al.*, 2021).

The literature consistently emphasizes that while the volume of publicly available omics data is growing rapidly, the quality and consistency of the associated metadata remain inadequate. Incomplete, inconsistent, or poorly standardized metadata hinders not only data reuse and integration but also compromises reproducibility. These challenges are especially critical in complex disease research, such as in RNA-seq studies of Amyotrophic Lateral Sclerosis (ALS).

Despite various efforts, current metadata submission systems often lack robust validation mechanisms and fail to enforce the use of controlled vocabularies or ontologies. As a result, datasets that might otherwise be highly valuable remain underutilized. Gonçalves and Musen (2019) demonstrate that metadata must not only exist but also be well-structured and ontology-driven to be effective. They argue for the necessity of metadata authoring tools that integrate standardization and enforce completeness at the point of submission, supporting the FAIR principles (Gonçalves and Musen, 2019). A defining attribute of FAIR metadata is its machine-readability, which enables computational systems to automatically find, access, interpret, and reuse data without human intervention. This requires the use of standardized formats, controlled vocabularies, and structured data models (Wilkinson *et al.*, 2016).

The literature increasingly emphasizes that a unified, FAIR-compliant metadata framework, supported by automated tools, clear standards, and community-wide adoption, is essential for fully utilizing the potential of RNA-seq and other high-throughput data.

1.3 Metadata Standards for RNA-seq

Creating FAIR-compliant data relies fundamentally on the completeness and quality of metadata. While access to raw data is essential, the accompanying metadata provides the context necessary for understanding, reproducing, and reusing that data (Leipzig *et al.*, 2021). This is especially true for RNA-seq, where high-throughput experiments generate complex datasets that vary widely depending on study design, sample origin, preparation methods, and sequencing platforms (Conesa *et al.*, 2016). Without detailed and standardized metadata, even the most well-produced RNA-seq data may become inaccessible or unusable over time.

Therefore, the scientific community has developed several metadata standards specifically aimed at capturing the necessary information to support reproducibility and reuse in transcriptomics research. The following subsections outline four of the most influential metadata standards in RNA-seq: Minimum Information About a Microarray Experiment (MIAME), Minimum Information about a High-Throughput Nucleotide Sequencing Experiment (MINSEQE), Functional Annotation of Animal Genomes (FAANG), and Human Cell Atlas (HCA).

1.3.1. MIAME guidelines

The original MIAME guidelines were published in 2001 by the Microarray Gene Expression Database (MGED) group (Brazma *et al.*, 2001). These guidelines defined the essential information that should accompany microarray data, which refers to gene expression data generated using DNA microarray technology, a method that measures the expression levels of thousands of genes simultaneously, to ensure reproducibility, data interpretation, and reuse (Wiltgen and Tilz, 2007). The initial MIAME guidelines were designed for microarray experiments but laid the groundwork for data standardization in high-throughput genomics and transcriptomics. MIAME provides a structured framework for reporting essential experimental details, which supports independent verification and improves the interpretability of results.

MIAME specifies six essential categories (Table 2) of information that should be included when reporting a microarray experiment.

Table 2

MIAME Metadata Categories (Brazma *et al.*, 2001)

Metadata Category	Description
Experimental design	The objectives, variables, and sample relationships.
Array design	Including the probe content, layout, and annotation.
Sample information	Source organism, treatments, and preparation.
Hybridization procedures	How labeled RNA was hybridized and washed.
Measurements	Scanner models, image formats, and extraction methods.
Normalization and data processing	The algorithms and parameters used.

Importantly, MIAME does not specify the format in which this information must be provided, but only its content. This flexibility helped increase adoption but also introduced challenges with consistency and interpretation across studies. To address this, a checklist (Appendix 1) was created and adopted by major journals, and data repositories such as ArrayExpress and Gene Expression Omnibus (GEO) began requiring MIAME-compliant submissions (Brazma, 2009). Edgar and Barrett (2006) reported that MIAME support in GEO facilitates independent evaluation and high repository usage, while Spellman et al. (2002) and Brazma (2009) showed that MIAME-compliant models enable standardized data exchange and clearer interpretation (Spellman *et al.*, 2002; Edgar and Barrett, 2006; Brazma, 2009).

At the same time, several studies have noted ongoing challenges. A study that reviewed data submission and MIAME compliance in 127 articles involving microarray-based microRNA profiling reported low or inconsistent compliance and a reluctance to share data,

factors that can compromise research quality (Witwer, 2013). Moreover, it has been challenging to find the optimal balance between submitter effort and the appropriate level of metadata detail to request, all within a rapidly evolving technological and social environment (Edgar and Barrett, 2006). MIAME has also received criticism for being insufficiently equipped to capture the full complexity of high-dimensional microarray data, with concerns that its minimal standards cannot account for unforeseen biological variables that may influence experimental outcomes (Galbraith, 2006).

These findings suggest that while MIAME guidelines enhance data sharing and validation, their effective implementation depends on managing technical complexity and ensuring consistent adherence.

1.3.2. MINSEQE guidelines

As next-generation sequencing technologies evolved and RNA-seq emerged as a more prominent tool in genomics, MIAME faced new challenges. In 2008, the Minimum Information about a High-throughput Nucleotide Sequencing Experiment (MINSEQE) standard was proposed (Brazma *et al.*, 2012). This version extended the principles of MIAME to accommodate the specifics of sequencing technologies, providing more detailed guidelines for RNA-seq, DNA sequencing, and other high-throughput sequencing platforms. The aim was to ensure that sequencing data are reusable, interpretable, and reproducible by defining a minimal set of metadata requirements.

MINSEQE specifies five essential categories (Table 3) of information that should accompany sequencing datasets (Brazma *et al.*, 2012).

Table 3

MINSEQE Metadata Categories (Brazma *et al.*, 2012)

Metadata Category	Description
Biological system, samples, and experimental variables	Describes the biological context of the study, including the organism, experimental design, grouping of samples, and variables under investigation (e.g. treatment, time point, genotype)
Sequence read data	Refers to the raw sequencing output, such as FASTQ files, and includes details about the sequencing platform, read length, and base-level quality scores.
Data processing	Covers the computational steps taken to process the raw reads, including alignment,

Metadata Category	Description
	quantification, normalization, software used, and relevant parameter settings.
General information and sample–data relationships	Provides metadata that links samples to data files, including unique identifiers and information about how biological samples correspond to sequencing libraries and outputs.
Experimental and data processing protocols	Documents the laboratory protocols used for library preparation, sequencing chemistry, and any other relevant methods. This section should also describe how the data were processed computationally, including pipeline versions and repository links.

Unlike MIAME, which focused on microarrays, MINSEQE was developed taking into account the complexity of sequencing pipelines, including raw data formats, computational workflows and repository submission practices (Brazma *et al.*, 2012).

Despite its conceptual value, there is a notable lack of published studies systematically evaluating the level of compliance with MINSEQE or assessing its impact on the quality and reusability of sequencing data. While MIAME has been the subject of multiple reviews and empirical assessments, MINSEQE remains largely unevaluated in practice, making it difficult to determine how consistently it is applied across public repositories or research publications (Brazma, 2009).

1.3.3. FAANG

In 2014, the FAANG Consortium was established to define and improve experimental, metadata and bioinformatics standards; ensure that experiments conducted to produce functional annotation adhere to these standards; and release all the experimental and metadata in an open access manner, rapidly and before publication (Andersson *et al.*, 2015). The aim of the Consortium is to produce comprehensive maps of functional elements in the genomes of domesticated animal species based on common standardized protocols and procedures (Andersson *et al.*, 2015). To support this, FAANG has introduced a structured set of metadata rule sets, including the sample core metadata rules, experiments core metadata rules, and analyses metadata rules (*FAANG*, 2025). These rule sets guide consistent metadata collection across different stages of the data lifecycle, ensuring interoperability and compliance with FAIR principles.

FAANG metadata guidelines are structured to be not only comprehensive but also practical and easy to implement. The consortium provides clear, rule-based templates for each

metadata category, detailing exactly what information is required, recommended, or optional, along with specific ontology terms and formatting instructions. For example, the metadata section for the "Organism" entity includes fields such as species (with NCBI Taxon ID), sex (linked to the PATO ontology), birth date, breed, and health status (Martínez-Romero *et al.*, 2017). Each field is annotated with its expected format, required ontology IDs, and validation rules, ensuring clarity and consistency across studies.

Despite its strengths, there are several limitations to using FAANG metadata guidelines for RNA-seq outside of agricultural genomics. FAANG was originally designed for domesticated animal species, and its terminology and structure reflect that focus. For example, fields such as "breed" or "production environment" may not translate meaningfully to human biomedical data or model organisms. Additionally, the framework was tailored more toward functional genome annotation than transcriptomic analysis, which means that technical RNA-seq metadata, such as quality control metrics, read length, sequencing depth, or single-cell parameters, are not explicitly covered.

Moreover, while FAANG promotes the use of standardized ontologies, the ontologies it prioritizes (e.g., LBO for breeds, PATO for phenotypes) may not fully align with those commonly used in human RNA-seq metadata annotation (Martínez-Romero *et al.*, 2017), such as The Human Disease Ontology (DO) for diseases or The Cell Ontology (CO) for cell types (Diehl *et al.*, 2016; Schriml *et al.*, 2019). Finally, FAANG's metadata schema may not integrate seamlessly with formats used by human-focused repositories like GEO or The Sequence Read Archive (SRA), which adhere to guidelines such as MIAME and MINSEQE.

Nonetheless, the clarity of FAANG's rules, the use of ontologies, and the structured documentation templates make it a valuable foundation that can be extended or adapted for broader RNA-seq metadata use cases.

To illustrate how FAANG supports metadata consistency for transcriptomic experiments, Table 4 outlines selected metadata fields from the FAANG experiment core metadata rules specifically relevant to RNA-seq. These fields are designed to capture key information about experimental design, protocols, and sequencing methodology. Although FAANG was originally developed for agricultural species, these fields offer a practical model for RNA-seq metadata in broader contexts. In addition to experiments, FAANG also defines structured metadata categories for samples and analyses, promoting consistency across the full data lifecycle, from biological source to computational output.

Table 4

FAANG Metadata Fields for RNA-seq Experiments (*FAANG*, 2025)

Field	Description
Experiment target	What the experiment was trying to find, list the text rather than an ontology link, for example 'polyA RNA'.
RNA preparation 3' adapter ligation protocol	Link to the protocol for 3’ adapter ligation used in preparation.
RNA preparation 5' adapter ligation protocol	Link to the protocol for 5’ adapter ligation used in preparation.
Library generation pcr product isolation protocol	Link to the protocol for isolating PCR products used for library generation.
Preparation reverse transcription protocol	Link to the protocol for reverse transcription used in preparation.
Library generation protocol	Link to the protocol used to generate the library.
Read strand	For strand specific protocol, specify which mate pair maps to the transcribed strand or Report 'non-stranded' if the protocol is not strand specific. For single-ended sequencing: use 'sense' if the reads should be on the same strand as the transcript, 'antisense' if on the opposite strand. For paired-end sequencing: 'mate 1 sense' if mate 1 should be on the same strand as the transcript, 'mate 2 sense' if mate 2 should be on the same strand as the transcript.
RNA purity 260:280 ratio	Sample purity assessed with fluorescence ratio at 260 and 280nm, informative for protein contamination.
RNA purity 260:230 ratio	Sample purity assessed with fluorescence ratio at 260 and 230nm, informative for contamination by phenolate ion, thiocyanates, and other organic compounds.
RNA integrity number	It is important to obtain this value, but if you are unable to supply this number (e.g. due to machine failure) then by submitting you are asserting the quality by visual inspection of traces and agreeing that the samples were suitable for sequencing.

1.3.4. HCA

The Human Cell Atlas (HCA) is an international collaborative initiative whose mission is to create comprehensive reference maps of all human cells, the fundamental units of life, as a basis for understanding human health and for diagnosing, monitoring, and treating disease (Regev *et al.*, 2017). HCA Metadata guidelines significantly enhance the standardization of RNA sequencing data by establishing a framework for comprehensive and highly detailed metadata documentation (Füllgrabe *et al.*, 2020). This standardization is crucial for ensuring reproducibility, facilitating data sharing, and improving the overall quality of scientific research. HCA metadata guidelines are structured around three complementary categories: Tier 1 metadata, Tier 2 metadata, and cell annotation metadata (*HCA Data Portal*, 2025). Tier 1 metadata provides the technical information necessary for data integration across studies. It

includes standardized, uniformly structured fields that help researchers assess and mitigate batch effects. This metadata is stored in the CellxGene Discover platform and formatted as AnnData matrices. In contrast, Tier 2 metadata captures biological information essential for downstream analysis and interpretation. This tier includes more detailed, often sensitive data that are protected through managed access and are tailored to the needs of individual Biological Networks. These data are made available through the HCA Data Repository and commonly stored in FASTQ format. Lastly, cell annotation metadata focuses on detailed cell-level descriptions such as cell type or state, and is captured primarily on a per-cell basis. This category supports biological labeling efforts and is maintained via the Cell Annotation Platform (CAP), also using AnnData format. Together, these structured tiers ensure that HCA metadata is both comprehensive and scalable for large-scale single-cell genomics research.

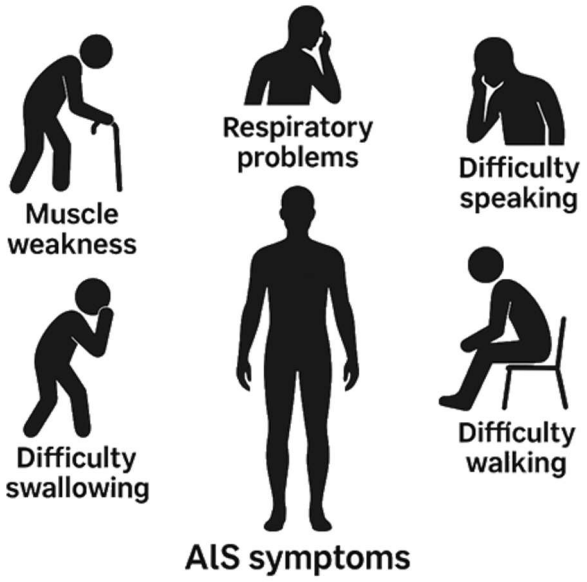
Despite its structured approach, the HCA metadata model faces several limitations. Tier 2 and cell-level metadata are not always consistently submitted or updated across studies (Lee *et al.*, 2025). Furthermore, the separation of metadata into distinct platforms may introduce fragmentation, making it harder for users to obtain a unified view of both technical and biological metadata.

1.4. Metadata in ALS Research

Amyotrophic Lateral Sclerosis (ALS) is a progressive and fatal neurodegenerative disorder that affects motor neurons in the brain and spinal cord (Feldman *et al.*, 2022). It is a clinically and genetically heterogeneous disease, with both sporadic and familial forms, and varying patterns of onset, progression, and molecular signatures (Zarei *et al.*, 2015). The figure 1 illustrates the clinical manifestations of ALS including muscle weakness, dysarthria, dysphagia, spasticity, and respiratory insufficiency, which result from the degeneration of both upper and lower motor neurons (Muneer *et al.*, 2014; Zarei *et al.*, 2015). As RNA sequencing becomes an increasingly important tool for uncovering transcriptomic alterations in ALS, the quality and completeness of associated metadata play a critical role in enabling robust interpretation, reproducibility, and secondary analysis (Kiaei and Kiaei, 2022).

Figure 1

Symptoms Associated with Amyotrophic Lateral Sclerosis (Zarei *et al.*, 2015)



Despite the growing availability of RNA-seq datasets related to ALS in public repositories, metadata reporting remains inconsistent (Cao and Scotter, 2022). In many cases, datasets include sequencing files with minimal or no contextual information, making it difficult to assess their relevance or quality for reuse.

The challenges of metadata collection in ALS are compounded by the disease's inherent variability (Swinnen and Robberecht, 2014). Biological samples may differ widely in terms of postmortem interval, clinical stage, or RNA integrity, all of which influence gene expression results. Moreover, the relatively small cohort sizes common in ALS studies further heighten the need for comprehensive metadata to support statistical power and meaningful stratification (van Rheenen *et al.*, 2021). Without well-documented metadata, it becomes nearly impossible to control for confounding variables or validate results in independent datasets.

2. Data and Methods

2.1 Development of a FAIR-Compliant RNA-seq Metadata Framework

This study was designed to evaluate existing metadata standards in RNA-seq research and to develop a structured, FAIR-compliant metadata framework. The methodology included a comprehensive literature review, a comparative analysis of existing metadata guidelines and the development of a metadata collection template using FAIR principles.

We identified and reviewed publicly available metadata standards and guidelines relevant to RNA-seq. The selection of frameworks was guided by the FAIR Cookbook Metadata Profile for Transcriptomics, which outlines essential metadata components to ensure compliance with FAIR data principles (Rocca-Serra *et al.*, 2023). The following frameworks were analyzed:

- MIAME - Minimum Information About a Microarray Experiment (Brazma *et al.*, 2001)
- MINSEQE - Minimum Information about a High-Throughput Nucleotide Sequencing Experiment (Brazma *et al.*, 2012)
- FAANG - Functional Annotation of Animal Genomes (Andersson *et al.*, 2015)
- HCA - Human Cell Atlas metadata guidelines (Regev *et al.*, 2017)

These standards were chosen because they are widely recognized in the community and collectively address a broad range of biological and technical variables relevant to transcriptomic research.

MIAME and MINSEQE are primarily presented as narrative guidelines that describe the types of metadata that should be reported. Since these standards do not include predefined templates, we manually derived structured templates from their descriptions. This involved interpreting the textual standards and translating them into standardized metadata fields organized into logical categories.

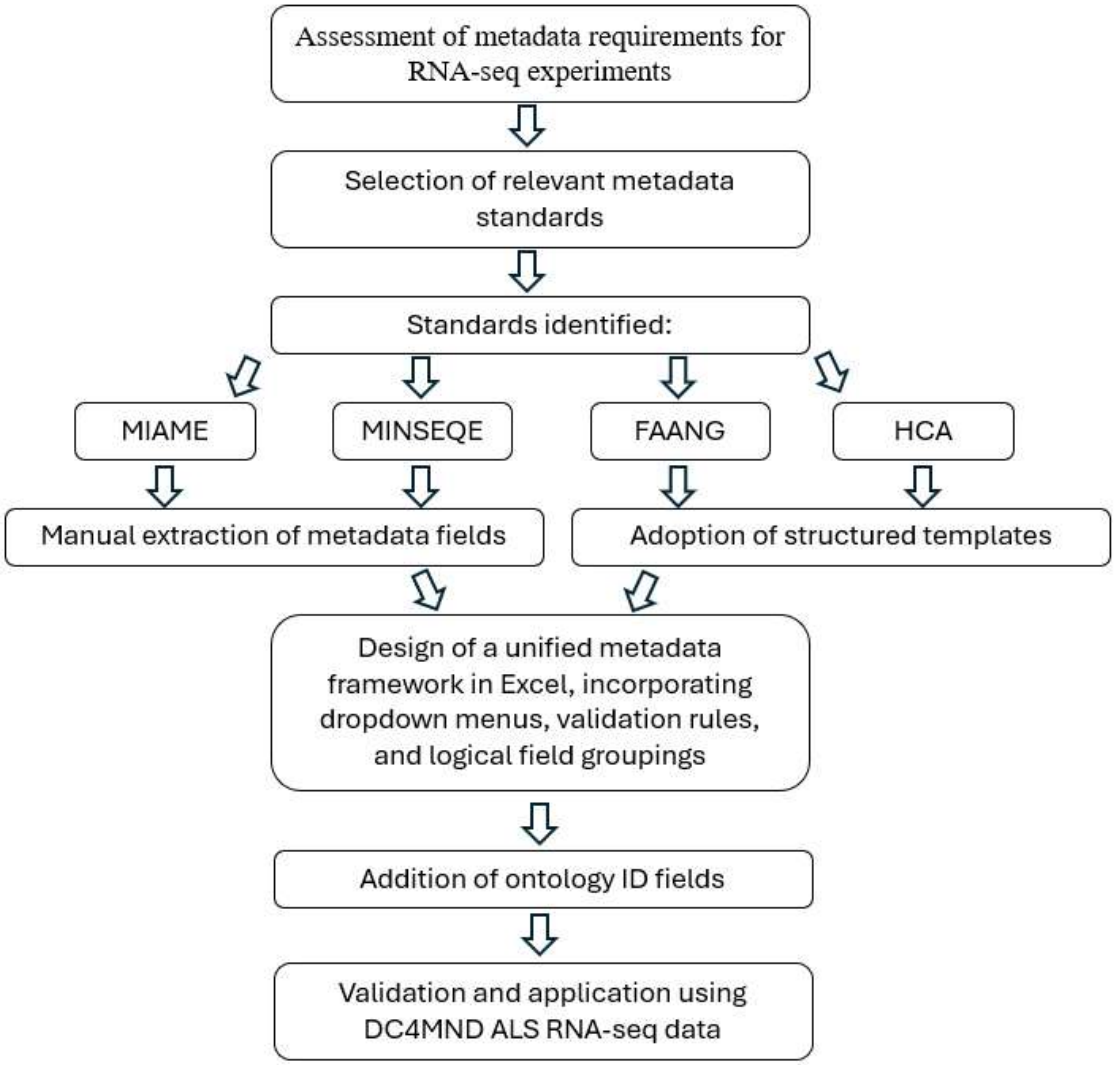
In contrast, FAANG and HCA already provide structured metadata templates with rule-based fields, validation logic, and controlled vocabulary support. These were used as-is, without modification, to preserve their original formatting and intended use.

To visually summarize the development process, a flowchart (Figure 2) illustrates the sequential steps involved in evaluating standards, extracting fields, and designing the unified metadata template. For each framework, a metadata template was designed using Microsoft Excel to support manual metadata entry, incorporating controlled vocabularies, drop-down menus for standardization, conditional formatting and cell validations. The structure supports

future translation into a user-friendly, automated, FAIR-compliant metadata entry and management system. Metadata templates based on MIAME, FAANG, and HCA guidelines are included in Appendices 2–4.

Figure 2

Flowchart of RNA-seq Metadata Framework Creation and Validation



2.2 MINSEQE-compliant metadata template

A complete MINSEQE-compliant metadata template, derived from the guideline's narrative description and expanded with ontology fields for enhanced FAIR compliance, is presented in this chapter. The following sections provide a structured overview of the metadata categories included in this framework. Each category is presented as a table detailing the

expected information, including a brief description, and answer type. To enhance readability and visual clarity, larger categories were divided into smaller, thematically grouped tables. Metadata fields were classified into common categories including:

1. Contact Information
2. Experimental Design
3. Sample Information
4. Sequencing and Data Processing

2.2.1. Contact Information

This section (Table 5) captures the essential identification and affiliation details of the data submitter. Including a unique researcher identifier, such as an ORCID ID, enhances transparency and enables clear attribution across research outputs (Shillum *et al.*, 2021). Accurate contact information also facilitates future communication and supports data trustworthiness.

Table 5

Contact Information

Field	Description	Answer Type	Link
Author (submitter)*	First and last name of the person submitting the data	Text	
Author`s ORCID	ORCID identifier providing a unique researcher ID	ORCID ID	https://orcid.org/
Email Address*	Primary email for correspondence	E-mail	
Institution Name*	Affiliated institution of the submitter	Text	
Contributors	Other individuals involved in the study	Text	
Links (URL)*	Link to related publication or data repository	Link	

*Required metadata field

2.2.2. Experimental Design

This section (Tables 6 and 7) outlines the core elements of the experimental design. It includes titles, descriptions, variables, replicates, and quality-control methods, all of which are essential for interpreting the results, assessing reproducibility, and guiding future experiments.

Table 6

Experimental Design. Experimental Overview

Field	Description	Answer Type
Experiment Title*	A brief, descriptive single-sentence title for the experiment.	Text
Description*	Free-text summary or a link to the experiment description/publication.	Text
Type of the Experiment*	E.g., normal vs. diseased, time course, dose response.	Text

*Required metadata field

Table 7

Experimental Design. Experimental Variables and Controls

Field	Description	Answer Type	Link
Experimental Factor	Key experimental variables (e.g., dose, time point)	Text	
Experimental Factor Value	Specific values for those factors (e.g., time duration)	Text	
Experimental Factor Units	Units used for the experimental factor values (e.g., hours, mg/mL).	Text	
Sample Treatment Type	Whether treatment was in vitro, in vivo, or not applied.	Multiple Choice	
Replicates	Usage and types of replicates	Multiple Choice	https://www.nature.com/documents/Biological_and_technical_replicates_guidelines.pdf
Quality-control	Quality control strategies	Text	https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8

2.2.3. Sample Information

This section captures essential metadata about the biological samples used in the study. Use of standardized ontologies (e.g., NCBI Taxonomy, Uberon Anatomy Ontology, Cell Ontology, Disease Ontology) is encouraged for consistent annotation of biological entities (Starr *et al.*, 2015; Diehl *et al.*, 2016; Schriml *et al.*, 2019; Schoch *et al.*, 2020). Table 8 focuses on sample identification and classification, while Table 6 details treatment and handling procedures.

Table 8

Sample Identification and Classification

Field	Description	Answer Type	Link
Sample ID	A unique identifier for each sample.	Text	
Organism	Species (e.g., Homo sapiens, Mus musculus).	Text	
Organism ID	Species name annotated using the NCBI Taxonomy.	Ontology ID	https://www.ncbi.nlm.nih.gov/taxonomy
Tissue Type	Tissue type.	Text	
Tissue Type ID	Tissue type annotated using the Uberon Anatomy Ontology	Ontology ID	https://www.ebi.ac.uk/ols/ontologies/uberon
Cell Type	Cell type.	Text	
Cell Type ID	Cell type annotated using the Cell Ontology.	Ontology ID	https://www.ebi.ac.uk/ols4/ontologies/cl
Qualifier	Defines a sample characteristic (e.g., strain, age, treatment).	Text	
Value	The value associated with the qualifier.	Text	
Value ID	The value ID	OntologyID	
Source	Source of information (e.g., DOID, lab records).	Text	https://www.disease-ontology.org/ https://www.ebi.ac.uk/ols4/ontologies/pato
Sample-Data Relationships	A table linking each sample to its corresponding raw and processed data files.	Text	

Table 9 addresses the experimental context of the samples, including treatments, preparation methods, and labeling.

Table 9

Sample Preparation and Treatment

Field	Description	Answer Type	Link
Sample Processing Protocols	How nucleic acids (RNA, DNA) were processed.	Text	
Extraction	Method used to extract nucleic acids.	Text	
Purification	Method for removing contaminants.	Text	
Preparation	Preparation steps (e.g., fragmentation, ligation).	Text	
Experimental Factors	Key experimental variables (e.g., drug dose, time point).	Text	

Field	Description	Answer Type	Link
Sample Treatment	Describes the treatment applied.	Text	
Sample Labeling	Labeling method for RNA or DNA.	Text	

2.2.4. Sequencing and Data Processing

This section provides metadata fields related to sequencing and data processing. To maintain readability, the section is organized into five structured tables: Sequence Read Data, Processed Data, Sequencing Methodology, Data Processing and Mapping, and Correction, Rejection, and Reference Information.

Table 10 captures raw sequencing outputs including base-level data, quality scores, and file formats.

Table 10

Sequence Read Data

Field	Description	Answer Type	Link
Raw Data	The raw sequencing data generated by the sequencing instrument.	FASTQ files	
Quality Scores	Base-level quality scores associated with each sequence read.	Text	
Quality Scores ID	A standardized ontology identifier that describes the quality scores, using terms from the EDAM Ontology	Ontology ID	https://edamontology.org
Scale	Description of the scale used for quality scores.	Text	
Data Format	The format of the raw sequencing data (e.g., FASTQ).	Multiple Choice	
Data Format ID	A standardized ontology identifier that describes the file format of the data, using terms from the EDAM Ontology	Ontology ID	https://edamontology.org

Table 11

Processed Data

Field	Description	Answer Type	Link
Processed Data	Final processed data used for analysis (e.g., gene expression matrix, peak calls).	File	
Data Format for Processed Data	Format for the processed data (e.g., CSV, Excel).	Text	
Data Format ID	A standardized ontology identifier that describes the file format of the data, using terms from the EDAM Ontology	Ontology ID	https://edamontology.org
Gene Expression Data Matrix	Matrix where each row is a gene and each column is a biological condition.	Spreadsheet	
Gene Expression Data Matrix ID	A standardized ontology identifier that describes the file format of the data, using terms from the EDAM Ontology	Ontology ID	https://edamontology.org

Table 12 provides details of the sequencing platform, preparation, and chemistry used.

Table 12

Sequencing Methodology

Field	Description	Answer Type
Sequencing Platform and Technology	Platform used for sequencing.	Text
Library Preparation Method	Method used to prepare RNA-seq libraries.	Text
Sequencing Chemistry	Sequencing chemistry/platform.	Text
Labelling Methodology	Labeling method (if any).	Text
Amplification Methodology	Method to amplify RNA/DNA.	Text

Table 13 captures tools and steps involved in aligning, filtering, and normalizing sequencing data.

Table 13

Data Processing and Mapping

Field	Description	Answer Type
Read Mapping and Alignment Methods	Tool and method used for alignment.	Text
Data Filtering and Quality Control Methods	Steps to clean/filter the data.	Text
Data Normalization Method	Normalization approach used.	Text
Data Analysis Tools and Algorithms	Tools for downstream analysis.	Text

Table 14 covers advanced correction, filtering, and reference genome identifiers.

Table 14

Correction, Rejection, and Reference Information

Field	Description	Answer Type
Data Rejection Methods	Method for excluding low-quality data.	Text
Data Correction Methods	Corrective methods for biases/batches.	Text
Alignment Methods	Tools used for alignment.	Text
Data Smoothing Methods	Techniques for smoothing count data.	Text
Data Filtering Methods	Filter criteria used.	Text
Identifiers	Reference genome version used.	Text

2.3 Case Study Application: ALS RNA-seq Data

To evaluate practical applicability, the template was tested using RNA-seq data from ALS research. Specifically, RNA-seq datasets from the DC4MND project were used to simulate real-world metadata curation. Metadata elements were sourced directly from participating scientists and partner institutions within the consortium. The completed metadata records were then evaluated for completeness, consistency, and alignment with the FAIR principles. This case study served to illustrate the framework's effectiveness in standardizing metadata and identifying common gaps that hinder data reuse in RNA-seq research.

2.4 Overview of the DC4MND Project

The DC4MND (Multidimensional Mechanistic Investigations of Trans-Spinal Direct Current Stimulation in Motor Neuron Disease; Nr. 7-3-2.N-1876, The EU Joint Programme – Neurodegenerative Disease Research (JPND)) project is a European, multi-institutional research initiative focused on exploring the mechanisms and therapeutic potential of trans-spinal direct current stimulation (tsDCS) in the context of ALS. The project brings together experts in electrophysiology, molecular biology, computational modeling, and bioinformatics to investigate how tsDCS influences motoneuron excitability, glial function, and spinal circuitry in both animal models and human iPSC-derived co-cultures. As part of its systems-level approach, the study incorporates RNA-seq analysis to assess transcriptomic changes in response to tsDCS treatment across different biological models.

In the animal model arm of the study, transcriptomic profiling was performed on spinal cord motor neurons isolated via laser capture microdissection from SOD1-G93A transgenic mice, one of the most widely used and well-characterized ALS models. These mice express a mutated form of the human superoxide dismutase 1 (SOD1) gene and exhibit hallmark ALS phenotypes, including progressive motor neuron degeneration, muscle weakness, and paralysis (Kim *et al.*, 2015). This model enables mechanistic insights into disease pathology and therapeutic responses.

The project also emphasizes rigorous data and metadata curation practices, aligning with FAIR principles to ensure reproducibility, transparency, and reusability of results.

2.5 Metadata Evaluation Using FAIR Metrics

To assess the completeness and FAIR compliance of the RNA-seq metadata produced using our framework, we applied a modified version of the FAIR assessment methodology developed by Elucidata (Chaudhr and Verma, 2021). This framework (Table 15) was originally designed to evaluate metadata quality in public transcriptomic repositories such as the GEO. It incorporates eight key metrics across the four FAIR principles, Findability, Accessibility, Interoperability, and Reusability. Each metric is scored based on the presence and richness of relevant metadata fields, such as the use of unique identifiers, standard ontologies, and machine-readable formats. Scores range from 0 (absent or insufficient information) to 1 (complete and compliant), with an overall composite FAIR score calculated as a weighted average. In our

adaptation, we manually evaluated each applicable field in our MINSEQE-compliant metadata template and calculated the resulting FAIR score accordingly.

Table 15

Elucidata's FAIR Assessment Metrics

FAIR Principle	Metric	Metric ID
F2. Data are described with rich metadata	Metadata contains biological keywords that describe the study	F2
F3. Metadata clearly and explicitly include the identifier of the data it describes	Metadata include machine actionable identifier to processed data	F3
I2. (meta)data use vocabularies that follow FAIR principles	Metadata are described by biological keywords and unique ontology identifiers	I2
R1.2. (meta)data are associated with detailed provenance	Metadata provide information regarding the overall design of the study	R1_2_1
	Metadata provide contact information of the creator(s)	R1_2_2
	Metadata specify the publication identifier	R1_2_3
	Metadata contain the platform ID	R1_2_4
	Processed data associated with the study contain sufficient processing context	R1_2_5

3. Results

To evaluate the applicability and completeness of the proposed metadata template, we applied it to a real-world RNA-seq dataset generated within the DC4MND project. The RNA-seq experiment involved spinal cord motor neurons isolated by laser capture microdissection from SOD1-G93A mice, a widely used transgenic model that mimics key features of human ALS pathology, treated with trans-spinal direct current stimulation (tsDCS), and sequenced using the Illumina NovaSeq 6000 platform.

3.1 DC4MND RNA-seq Metadata

3.1.1. DC4MND RNA-seq Metadata: Contact Information

The Contact Information section (Table 16) of the metadata was the most straightforward to complete. All required fields were filled without difficulty, the only missing element was the URL linking to the dataset, which will be added once the RNA-seq data is publicly deposited. While RNA-seq data in the DC4MND project involved contributions from multiple researchers and collaborators, this thesis provides the metadata details of the primary submitter only. A full list of co-authors and contributors will be available in the final data repository and associated publications.

Table 16

DC4MND RNA-seq Metadata. Contact Information.

Field	Description	Value
Author (submitter)	First and last name of the person submitting the data	Edīte Vārtiņa
Email Address	Primary email for correspondence	Edite.Vartina@rsu.lv
Author`s ORCID	ORCID identifier providing a unique researcher ID	https://orcid.org/0000-0002-9639-2883
Institution Name	Affiliated institution of the submitter	Riga Stradins University, Bioinformatics Group
Contributor 1	Other individuals involved in the study	Baiba Vilne
Email Address	Primary email for correspondence	Baiba Vilne@rsu.lv
ORCID	ORCID identifier providing a unique researcher ID	https://orcid.org/0000-0002-1084-7067
Institution Name	Affiliated institution of the submitter	Riga Stradins University, Bioinformatics Group

Field	Description	Value
Contributor 2	Other individuals involved in the study	Tatjana Kiseļova
Email Address	Primary email for correspondence	Tatjana.Kiselova@rsu.lv
ORCID	ORCID identifier providing a unique researcher ID	https://orcid.org/0009-0003-3853-3056
Institution Name	Affiliated institution of the submitter	Riga Stradins University, Bioinformatics Group
Contributor 3	Other individuals involved in the study	Egija Berga-Švītiņa
Email Address	Primary email for correspondence	Egija.Berga-Svitina@rsu.lv
ORCID	ORCID identifier providing a unique researcher ID	https://orcid.org/0000-0001-5150-0185
Institution Name	Affiliated institution of the submitter	Riga Stradins University, Bioinformatics Group
Contributor 4	Other individuals involved in the study	Burak Özkan
Email Address	Primary email for correspondence	burak.oezkan@uni-ulm.de
ORCID	ORCID identifier providing a unique researcher ID	https://orcid.org/0000-0002-5228-4880
Institution Name	Affiliated institution of the submitter	Ulm University
Contributor 5	Other individuals involved in the study	Francesco Roselli
Email Address	Primary email for correspondence	francesco.roselli@uni-ulm.de
ORCID	ORCID identifier providing a unique researcher ID	https://orcid.org/0000-0001-9935-6899
Institution Name	Affiliated institution of the submitter	Ulm University
Contributor 6	Other individuals involved in the study	Marcin Bączyk
Email Address	Primary email for correspondence	baczyk@awf.poznan.pl
ORCID	ORCID identifier providing a unique researcher ID	https://orcid.org/0000-0002-2656-9655
Institution Name	Affiliated institution of the submitter	Poznań University of Physical Education
Links (URL)	Link to related publication or data repository	

3.1.2. DC4MND RNA-seq Metadata: Experimental Design

The following tables present the completed metadata fields related to the experimental design of the study. Table 17 provides an overview of the experiment and context within the broader DC4MND project. It captures key details such as the experimental type and a descriptive summary of the biological questions being addressed.

Table 18 outlines the experimental variables and control conditions, specifying the key factors under investigation, acute and chronic tsDCS. These entries were completed based on experimental design documentation and the planned RNA-seq sampling strategy.

Table 17

DC4MND RNA-seq Metadata. Experimental Overview.

Field	Description	Value
Experiment Title	A brief, descriptive single-sentence title for the experiment.	Multidimensional mechanistic investigations of trans spinal direct current stimulation in motor neuron disease
Description	Free-text summary or a link to the experiment description/publication.	This metadata describes the RNA-seq component of the DC4MND project, which investigates the transcriptomic effects of trans-spinal direct current stimulation (tsDCS) in the context of Amyotrophic Lateral Sclerosis (ALS). The experiment examines both acute and chronic tsDCS effects on spinal motor circuits using ALS mouse models. This RNA-seq analysis was performed to assess gene expression changes in spinal cord motoneurons of SOD1-G93A mice in response to tsDCS stimulation, compared to sham-treated controls. https://neurodegenerationresearch.eu/wp-content/uploads/2023/05/DC4MND-Project-Fact-Sheet.pdf
Type of the Experiment	E.g., normal vs. diseased, time course, dose response.	Comparative analysis of tsDCS effects in diseased (ALS) vs. control animal models using time course response assessments.

Table 18

DC4MND RNA-seq Metadata. Experimental Variables and Controls

Field	Description	Value	
Experimental Factor	Key experimental variables (e.g., dose, time point)	Acute tsDCS	Chronic tsDCS
Experimental Factor Value	Specific values for those factors (e.g., time duration)	20 minutes after stimulation	One day after stimulation
Experimental Factor Units	Units used for the experimental factor values (e.g., hours, mg/mL).	minutes	days

3.1.3. DC4MND RNA-seq Metadata: Sample Information

The Sample Identification and Classification metadata (Table 19) provides detailed information on each biological sample, with each representing an individual mouse included in the RNA-seq experiment. Each sample is uniquely identified and annotated using standardized ontology references (e.g., NCBITaxon, UBERON, CL, MGI). Notably, all samples were derived from Mus musculus (NCBITaxon:10090), and motoneurons were isolated from the spinal cord.

It is also important to note that two different naming conventions were used for the RNA-seq samples across project documents. In experimental records, samples were labeled as S1-3 (sham/control group) and A3-5 (tsDCS stimulation group), while in the sequencing output and data files, samples appear under technical identifiers such as S11596Nr1 and S11596Nr6. The metadata table explicitly links both naming formats and connects each sample to its corresponding raw and processed data files to ensure clarity and traceability.

Table 19

DC4MND RNA-seq Metadata. Sample Identification and Classification

Field	Description	Value
Sample ID	A unique identifier for each sample.	S1-2 (S11596Nr1) - Sham -control S2-3(S11596Nr2) - Sham -control S3-3(S11596Nr3) - Sham -control A1-3 (S11596Nr4) - tsDCS-treated mice A 3-3 (S11596Nr6) - tsDCS-treated mice A4-4 (S11596Nr7) - tsDCS-treated mice A5-2 (S11596Nr8) - tsDCS-treated mice
Organism	Species (e.g., Homo sapiens, Mus musculus).	Mus musculus
Organism ID	Species name annotated using the NCBI Taxonomy.	NCBITaxon:10090
Tissue/Cell Type	Tissue or cell type.	Spinal cord (UBERON:0002240) Motorneuron (CL:0000100)
Tissue/Cell Type ID	Tissue or cell type. Uberon Anatomy Ontology or Cell Ontology.	UBERON:0002240 CL:0000100
Qualifier	Defines a sample characteristic (e.g., strain, age, treatment).	Genotype
Value	The value associated with the qualifier.	SOD1-G93A
Value ID	The value ID	MGI:2183719

Field	Description	Value	
Source	Source of information (e.g., DOID, lab records).	MGI database	
Sample-Data Relationships	A table linking each sample to its corresponding raw and processed data files.	S1-2 (S11596Nr1)	RNA_S11596Nr1.1.fastq.gz RNA_S11596Nr1.2.fastq.gz RNA_S11596Nr1.bam RNA_S11596Nr1.Aligned.sortedByCoord.out.bam

Table 20 continues with metadata related to sample preparation and treatment protocols, including RNA extraction, purification, and labeling. All procedures were carried out using the QIAGEN RNeasy Micro Kit (QIAGEN, Hilden, Germany), according to the manufacturer's instructions, and the RNA samples underwent fragmentation and adapter ligation steps in preparation for sequencing. The samples represent both tsDCS-treated and sham-treated conditions, with treatment performed in vivo. No additional labeling steps were applied, as the RNA-seq protocol involved direct sequencing of prepared cDNA libraries.

Table 20

DC4MND RNA-seq Metadata. Sample Preparation and Treatment

Field	Description	Value
Sample Processing Protocols	How nucleic acids (RNA, DNA) were processed.	RNA extracted using QIAGEN RNeasy Micro Kit
Extraction	Method used to extract nucleic acids.	QIAGEN RNeasy Micro Kit
Purification	Method for removing contaminants.	RNeasy spin columns
Preparation	Preparation steps (e.g., fragmentation, ligation).	RNA fragmentation and adapter ligation
Experimental Factors	Key experimental variables (e.g., drug dose, time point).	tsDCS treatment, genotype
Sample Treatment Type	Whether treatment was in vitro, in vivo, or not applied.	in vivo
Sample Labeling	Labeling method for RNA or DNA.	No labelling; direct RNA sequencing

3.1.4. DC4MND RNA-seq Metadata: Sequencing and Data Processing

The metadata fields related to sequencing and data processing are presented in Tables 21 to 23. These entries cover the entire technical workflow, from raw read generation and quality assessment to alignment, normalization, and downstream analysis. Table 20 outlines the basic

characteristics of the sequence read data, including file formats, quality score encoding, and data standards (e.g., EDAM ontology identifiers).

Table 21

DC4MND RNA-seq Metadata. Sequence Read Data

Field	Description	Value
Raw Data	The raw sequencing data generated by the sequencing instrument.	FASTQ files from Novaseq 6000 platform (Illumina, San Diego, USA)
Quality Scores	Base-level quality scores associated with each sequence read.	Phred quality scores, assessed using FastQC v0.11.9
Scale	Description of the scale used for quality scores.	Phred+64 (Sanger / Illumina 1.9 encoding)
Data Format	The format of the raw sequencing data (e.g., FASTQ).	FASTQ
Data Format ID	A standardized ontology identifier from the EDAM Ontology	EDAM:format_1930

Table 22 focuses on the processed data output, specifically the TMM (Trimmed Mean of M-values) normalized gene expression matrix, which forms the basis for biological interpretation and statistical analysis. However, some fields in this section are currently not filled in, as the final steps of the data processing workflow have not yet been completed. These metadata fields will be updated as soon as the processed data becomes available.

Table 22

DC4MND RNA-seq Metadata. Processed Data

Field	Description	Value
Processed Data	Final processed data used for analysis (e.g., gene expression matrix, peak calls).	TMM normalised gene expression matrix
Data Format for Processed Data	Format for the processed data (e.g., CSV, Excel).	CSV/TSV
Data Format ID	A standardized ontology identifier that describes the file format of the data, using terms from the EDAM Ontology	EDAM:format_3752 EDAM:format_3475
Gene Expression Data Matrix	Matrix where each row is a gene and each column is a biological condition.	

Field	Description	Value
Gene Expression Data Matrix ID	A standardized ontology identifier that describes the file format of the data, using terms from the EDAM Ontology	

Table 23 provides a comprehensive account of the sequencing methodology. It is important to note that this section is the most challenging to complete accurately without deep familiarity with the data processing pipeline. Much of this information resides in specialized protocols, sequencing facility documentation, or embedded within software pipelines, and is not immediately visible to external reviewers.

Table 23

DC4MND RNA-seq Metadata. Sequencing Methodology

Field	Description	Value
Sequencing Platform and Technology	Platform used for sequencing.	Novaseq 6000 platform (Illumina, San Diego, USA)
Library Preparation Method	Method used to prepare RNA-seq libraries.	SMART-Seq Stranded (Takara Bio)
Sequencing Chemistry	Sequencing chemistry/platform.	
Labelling Methodology	Labeling method (if any).	
Amplification Methodology	Method to amplify RNA/DNA.	
Read Trimming (if applicable)	Tools (their versions and parameters) to remove adapter sequences and low-quality bases (e.g. Trimmomatic, Cutadapt)	Adapters trimmed with Skewer v0.2.2; quality trimming not performed
Read Mapping and (Pseudo)Alignment Methods	Tools (their versions and parameters) used to map and align reads to the reference genome (e.g. STAR, HISAT2) or perfrom pseudoalignment transcriptome (Salmon, Kallisto).	Trimmed raw reads were aligned to mm10 using STAR (version 2.7.3)
Reference genome ID	Identifiers used for reference genomes	Mus_musculus.GRCm39.112.gtf
Read Mapping Quality Control Methods	Tools (their versions and parameters) that evaluate the quality and integrity of	FastQC v0.11.9

	aligned sequencing reads by analyzing coverage, biases, and mapping statistics, especially for RNA-seq data. (e.g., QualiMap).	
Quantification and Annotation	Tools (their versions and parameters) (e.g. featureCounts) to quantifying gene expression by counting how many aligned RNA-seq reads overlap with genomic features defined in the annotation file (GTF or GFF) and the annotation file (GENCODE, Ensembl) used.	Rsubread v1.32.4 using featureCounts
Data Filtering	Low count filtering to remove low-expressed genes and reduce noise in downstream statistical analysis.	Custom Python script was used to filter out genes that did not show a median expression ≥ 10 reads in at least one experimental group.
Data Transformation	To stabilize variance and normalize count distributions before visualization or clustering.	Log2 transformation
Data Normalization Method	Methods (tools, their versions and parameters) used to statistically normalize the data (TMM, Upper Quartile, DESeq2 Size Factor, Quantile Normalization) or expression normalization metrics used (e.g., TPM, RPKM, FPKM).	TMM, as implemented in DESeq2 (version 1.24.0) in R (version 4.0.4)
Downstream Data Analysis Tools and Algorithms	Tools (their versions and parameters) used for downstream data analysis (e.g., DESeq2, LIMMA for differential expression).	DESeq2 (version 1.24.0) for differential expression

3.2 Controlled Field Validation Log

Following the initial metadata curation, all fields that required controlled vocabulary terms or ontology identifiers were systematically reviewed. Validation was performed manually against established resources, including:

- NCBI Taxonomy for organism identifiers (Mus musculus → NCBITaxon:10090)
- UBERON for tissue annotations (spinal cord → UBERON:0002240)
- Cell Ontology for tissue and cell type annotations (motoneuron → CL:0000100)

- EDAM Ontology for data formats (FASTQ → EDAM:format_1930; CSV → EDAM:format_3752)
- MGI Database for genotype information (SOD1-G93A → MGI:2183719).

No invalid terms were identified during this validation step. Incomplete fields, such as dataset URLs and final annotation file versions, were marked as placeholders to be completed once data submission is finalized.

Table 24 summarizes the evaluation of the DC4MND RNA-seq metadata using FAIR assessment metrics adapted from Elucidata's framework for GEO datasets. Based on this evaluation, the composite FAIR score of the DC4MND RNA-seq metadata was 80.0 out of 100, indicating a high level of FAIR compliance. This score reflects particularly strong performance in metadata richness, the use of machine-readable identifiers, and provenance documentation. The result demonstrates the effectiveness of the structured framework in producing reusable and transparent metadata suitable for large-scale transcriptomic data sharing and integration.

Table 24

Evaluation of DC4MND Metadata Using Elucidata's FAIR Assessment Metrics

FAIR Principle	Metric	Assessment	Evidence
F2	Metadata contains biological keywords that describe the study	Yes	Includes keywords like 'Amyotrophic Lateral Sclerosis', 'tsDCS', 'motoneurons', 'SOD1-G93A'
F3	Metadata include machine actionable identifier to processed data	Yes	Processed data files (.bam, .fastq.gz) are linked to sample IDs and traceable
I2	Metadata are described by biological keywords and unique ontology identifiers	Yes	Uses NCBITaxon:10090, UBERON:0002240, CL:0000100
R1_2_1	Metadata provide information regarding the overall design of the study	Yes	Experimental design and variable tables included with timepoints and treatment details
R1_2_2	Metadata provide contact information of the creator(s)	Yes	Submitter name, email, ORCID, and institution are listed

FAIR Principle	Metric	Assessment	Evidence
R1_2_3	Metadata specify the publication identifier	Partially	Only project summary link provided, no DOI or PubMed ID yet
R1_2_4	Metadata contain the platform ID	Yes	Platform 'NovaSeq 6000' and tools like STAR, DESeq2 are documented
R1_2_5	Processed data associated with the study contain sufficient processing context	Yes	Detailed info on trimming, alignment, quantification, normalization methods

4. Discussion

The ongoing reproducibility "crisis" in biomedical research continues to raise major challenges for scientific progress and public credibility, as confirmed by multiple international surveys. In a 2016 survey conducted by Nature, more than 70% of researchers reported they had been unable to reproduce another scientist's experiment, and over half said they had failed to reproduce their own results (Baker, 2016). In the same study, over half of the respondents (52%) considered this a significant crisis, although many still expressed confidence in the overall reliability of the literature. Similarly, a cross-sectional survey published in 2024, which focused specifically on biomedical researchers found that 72% believed a crisis exists, and only 16% reported their institutions had established procedures to enhance reproducibility (Cobey *et al.*, 2024).

One of the widely endorsed solutions to this crisis is the adoption of the FAIR data principles - Findability, Accessibility, Interoperability, and Reusability (Wilkinson *et al.*, 2016). These principles emphasize not just open data, but usable data (Cobey *et al.*, 2024). In the context of RNA-seq, this means ensuring that metadata is complete, structured, and standardized (Caliskan, Dangwal and Dandekar, 2023). Such metadata allows researchers to assess the relevance and quality of public datasets, integrate them with new data, and perform meta-analyses. Incomplete or inconsistent metadata, by contrast, severely limits the utility of even high-quality sequencing data (Sabot, 2022). It can lead to misinterpretation, inappropriate comparisons, or the outright exclusion of otherwise valuable datasets from integrative studies (Glasziou *et al.*, 2014). Thus, improving metadata completeness is a necessary step toward unlocking the full potential of publicly available transcriptomic resources.

Our proposed framework is designed to systematically address persistent challenges in RNA-seq metadata curation, including inconsistency, incompleteness, and lack of standardization. To build this framework, we compiled and synthesized four widely used RNA-seq metadata standards to construct a unified template. This template is designed to prompt researchers to provide comprehensive metadata that covers essential biological and technical parameters. By drawing on established guidelines, we ensured that the fields included are both relevant to RNA-seq workflows and sufficient to support reproducibility and reuse.

One of the major limitations in current RNA-seq metadata practices is the frequent fragmentation and incompleteness of essential information. Essential biological and technical information, such as tissue origin, disease state, library preparation protocols, or sequencing settings, is often missing or only partially described (Sabot, 2022). This lack of detail can make

even high-quality sequencing data difficult to interpret or reuse. To address this, our framework emphasizes the identification and inclusion of mandatory metadata fields that are essential for interpreting and reusing the data. These core fields serve not only as documentation but also as indicators of dataset completeness, helping researchers quickly evaluate whether a dataset is sufficiently annotated for their intended analyses. Essential elements include clear sample-to-data relationships (e.g., linking individual biological samples to corresponding FASTQ or BAM files), biological metadata (such as tissue type, cell line, disease status, developmental stage, or treatment condition), and technical metadata (including library preparation method, sequencing platform, read layout, and strandedness) (Brazma *et al.*, 2012). By clearly defining which metadata elements are required, the framework encourages consistent reporting and facilitates data integration, ultimately supporting more reproducible and reusable transcriptomic research. One of the primary challenges with RNA-seq data is its high dimensionality. A typical RNA-seq dataset contains expressions for thousands of genes across hundreds or thousands of individual cells or samples. This large volume of data can be difficult to analyze and interpret without detailed metadata that describes not just the data itself but the experimental conditions that affect gene expression. For example, the same gene may exhibit varying expression levels depending on the tissue, disease state, or time of collection, all of which need to be captured in the metadata. This challenge is further compounded in large-scale, multi-center projects where different parts of the experiment may be conducted at different sites, often introducing variation in protocols, instrumentation, or data handling practices (Ulrich *et al.*, 2022). Such complexity is particularly pronounced in studies of heterogeneous and multifactorial diseases like amyotrophic lateral sclerosis, where capturing detailed, standardized metadata across sites is essential for meaningful analysis and comparison (Chiò *et al.*, 2013). Another challenge in RNA-seq research is the variability in experimental protocols. RNA-seq protocols can differ widely in terms of the methods used for sample collection, RNA extraction, library preparation, sequencing platforms, and data analysis pipelines (Abdelaziz *et al.*, 2024). For example, different library preparation methods or sequencing platforms can yield datasets with distinct biases and characteristics. These variations in methodology must be carefully documented in the metadata to ensure that the data can be properly interpreted, compared, and integrated.

To address this, our framework includes a guided metadata template featuring fill-in fields and multiple-choice options. This structure helps researchers navigate which data are required, minimizes ambiguity, and promotes consistency across contributors and study sites. By simplifying the process of metadata collection, the template supports more complete and harmonized documentation, even in complex, distributed research settings. One important

consideration in this process is the section covering sequencing methodology. This part of the metadata is often the most difficult to complete accurately without domain-specific knowledge of the data processing pipeline. Key details, such as sequencing chemistry, reference annotation sources, or quantification tools, are frequently embedded in specialized protocols, facility documentation, or within software environments and may not be apparent to external reviewers. For this reason, we recommend that these fields be completed by the researcher directly involved in the sequencing or bioinformatic analysis. Their expertise is essential to ensure that the metadata is both accurate and complete, which is particularly important for interpreting technical variability across datasets.

In addition to consistency and completeness, standardization is a fundamental goal of our framework. Free-text fields and unstructured inputs are common in metadata submissions, but they lack semantic clarity and hinder automation. By incorporating structured ontologies and controlled vocabularies, such as the Uberon anatomy ontology for tissue and organ terms, the NCBI Taxonomy for organism classification, and the Cell Ontology for cell-type classification, our framework enables machine-readable annotations and semantic interoperability (Diehl *et al.*, 2016; Schoch *et al.*, 2020). These features not only make the data more accessible to computational tools but also ensure alignment with community-driven standards, increasing the chances of long-term reusability.

To quantitatively assess the FAIRness of our proposed metadata framework, we applied Elucidata's FAIR assessment framework, which evaluates compliance with core FAIR principles through eight weighted metrics (Chaudhr and Verma, 2021). Based on this evaluation, the metadata generated for the DC4MND RNA-seq dataset received a composite FAIR score of 80 out of 100. This high score reflects strong performance across areas such as metadata richness, provenance, identifier usage, and interoperability. Compared to average scores observed in uncurated GEO datasets, this result demonstrates the value of prospective metadata design and highlights the framework's effectiveness in producing machine-actionable and reusable transcriptomic metadata (Chaudhr and Verma, 2021) .

Several previous efforts have aimed to improve RNA-seq metadata through retrospective curation or templated annotation tools. For example, MetaSRA focuses on enriching existing metadata from public repositories by mapping free-text descriptions to ontology terms, thereby improving semantic structure and comparability across studies. It uses a schema inspired by the ENCODE project and combines automated entity recognition with ontology-based normalization to label human RNA-seq samples from the Sequence Read Archive (SRA) (Bernstein, Doan and Dewey, 2017). This includes mapping to key biomedical

ontologies such as the Disease Ontology, Cell Ontology, and Uberon, along with extracting real-value properties like age or passage number. The MetaSRA project has since expanded to include annotated metadata for both human and mouse samples, broadening its relevance for comparative and cross-species studies. However, because MetaSRA operates retroactively, it cannot address gaps in original metadata completeness or inaccuracies present at the time of submission.

MetaSheet offers spreadsheet-based templates for metadata annotation with an emphasis on ease of use, allowing users to generate documentation and code from Excel or Google Sheets without requiring programming experience (Metadata Technology North America, 2025). It is particularly designed to support the GO FAIR initiative by enabling rapid metadata capture for statistical and scientific datasets. However, despite its accessibility and flexibility, MetaSheet lacks integration with controlled vocabularies or domain-specific ontologies, and does not currently support automated validation of metadata fields. Additionally, there is no RNA-seq–specific template available, and its general-purpose design may limit its immediate applicability for complex omics data.

MetaRNA-Seq provides a web-based tool for browsing, searching, and annotating RNA-seq metadata with study-level summaries and a semiautomated curation interface (Kumar *et al.*, 2015). Unlike the MetaSRA, which emphasizes experiment-level information, MetaRNA-Seq organizes metadata hierarchically and provides study-wide overviews, allowing users to filter by attributes such as disease type, replicate type, or time series design. The tool also offers guided annotation suggestions based on regular expression matches across biosamples and experiments. However, MetaRNA-Seq currently focuses only on human datasets and relies heavily on manual curation by a small number of contributors, limiting scalability and consistency across large repositories.

In contrast, our framework is designed to be implemented prospectively, at the point of data generation. This approach promotes higher metadata quality from the outset and reduces the need for retrospective correction. Additionally, our use of structured ontologies, a guided template with required fields, and automation-ready formatting provides a more robust and reusable solution. A current limitation, however, is the need for manual input from domain experts during the annotation process, particularly for technically detailed fields, highlighting a potential area for future automation and integration with sequencing pipelines.

Conclusions

The ongoing reproducibility "crisis" in biomedical research highlights the urgent need for improved data practices, particularly in the handling of metadata. In RNA-seq research, where high-dimensional datasets and complex experimental designs are common, incomplete or inconsistent metadata remains a major barrier to data reuse, integration, and interpretation. Our work addresses these challenges by presenting a structured framework for RNA-seq metadata curation that prioritizes completeness, consistency, and standardization.

By synthesizing four established RNA-seq metadata standards, we developed a unified template that guides researchers through the metadata annotation process using fill-in fields and multiple-choice options. This approach helps ensure the inclusion of critical biological and technical parameters, supports reproducibility, and aligns with FAIR data principles. We further emphasize the importance of domain expertise in accurately completing sections related to sequencing methodology and analysis pipelines, areas that are often difficult to annotate retrospectively.

References

1. Abdelaziz, E. H., Ismail, R., Mabrouk, M. S., & Amin, E. (2024). Multi-omics data integration and analysis pipeline for precision medicine: Systematic review. *Computational Biology and Chemistry*, *113*, 108254. https://doi.org/10.1016/j.compbiolchem.2024.108254
2. Andersson, L., Archibald, A. L., Bottema, C. D., Brauning, R., Burgess, S. C., Burt, D. W., Casas, E., Cheng, H. H., Clarke, L., Couldrey, C., Dalrymple, B. P., Elsik, C. G., Foissac, S., Giuffra, E., Groenen, M. A., Hayes, B. J., Huang, L. S. S., Khatib, H., Kijas, J. W., … Zhou, H. (2015). Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biology*, *16*(1), 4–9. https://doi.org/10.1186/s13059-015-0622-4
3. Baker, M. (2016). IS THERE A REPRODUCIBILITY CRISIS? *Nature*, *533*(7604), 452–454.
4. Baxi, E. G., Thompson, T., Li, J., Kaye, J. A., Lim, R. G., Wu, J., Ramamoorthy, D., Lima, L., Vaibhav, V., Matlock, A., Frank, A., Coyne, A. N., Landin, B., Ornelas, L., Mosmiller, E., Thrower, S., Farr, S. M., Panther, L., Gomez, E., … Rothstein, J. D. (2022). Answer ALS, a large-scale resource for sporadic and familial ALS combining clinical and multi-omics data from induced pluripotent cell lines. *Nature Neuroscience*, *25*(2), 226–237. https://doi.org/10.1038/s41593-021-01006-0
5. Bernstein, M. N., Doan, A., & Dewey, C. N. (2017). MetaSRA: Normalized human sample-specific metadata for the Sequence Read Archive. *Bioinformatics*, *33*(18), 2914–2923. https://doi.org/10.1093/bioinformatics/btx334
6. Bhandary, P., Seetharam, A. S., Arendsee, Z. W., Hur, M., & Wurtele, E. S. (2018). Raising orphans from a metadata morass: A researcher's guide to re-use of public 'omics data. *Plant Science*, *267*(November 2017), 32–47. https://doi.org/10.1016/j.plantsci.2017.10.014
7. Brazma, A. (2009). Minimum information about a microarray experiment (MIAME) - Successes, failures, challenges. *TheScientificWorldJournal*, *9*, 420–423. https://doi.org/10.1100/tsw.2009.57
8. Brazma, A., Ball, C., Bumgarner, R., Furlanello, C., Miller, M., Quackenbush, J., Reich, M., Rustici, G., Stoeckert, C., Trutane, S. C., & Taylor, R. (2012). MINSEQE: Minimum Information about a high-throughput Nucleotide SeQuencing Experiment - a proposal for standards in functional genomic data reporting. *Zenodo*, *0*. https://doi.org/10.5281/zenodo.5706411
9. 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., … Vingron, M. (2001). Minimum information about a microarray experiment (MIAME) - Toward standards for microarray data. *Nature Genetics*, *29*(4), 365–371. https://doi.org/10.1038/ng1201-365
10. Caliskan, A., Dangwal, S., & Dandekar, T. (2023). Metadata integrity in bioinformatics: Bridging the gap between data and knowledge. *Computational and Structural Biotechnology Journal*, *21*(August), 4895–4913. https://doi.org/10.1016/j.csbj.2023.10.006
11. Cao, M. C., & Scotter, E. L. (2022). Transcriptional targets of amyotrophic lateral sclerosis/frontotemporal dementia protein TDP-43 - meta-analysis and interactive graphical database. *Disease Models & Mechanisms*, *15*(9). https://doi.org/10.1242/dmm.049418
12. Chaudhr, M., & Verma, P. (2021). *FAIR annotation and evaluation of RNA sequencing and microarray data – Elucidata*. https://fairtoolkit.pistoiaalliance.org/use-cases/fair-annotation-and-evaluation-of-rna-sequencing-and-microarray-data-elucidata/
13. Chiò, A., Logroscino, G., Traynor, B. J., Collins, J., Simeone, J. C., Goldstein, L. A., & White, L. A. (2013). Global epidemiology of amyotrophic lateral sclerosis: a systematic review of the published literature. *Neuroepidemiology*, *41*(2), 118–130. https://doi.org/10.1159/000351153

14. Cobey, K. D., Ebrahimzadeh, S., Page, M. J., Thibault, R. T., Nguyen, P.-Y., Abu-Dalfa, F., & Moher, D. (2024). Biomedical researchers' perspectives on the reproducibility of research. *PLoS Biology*, *22*(11), e3002870. https://doi.org/10.1371/journal.pbio.3002870

15. Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*, 13. https://doi.org/10.1186/s13059-016-0881-8

16. Deshpande, D., Chhugani, K., Chang, Y., Karlsberg, A., Loeffler, C., Zhang, J., Muszyńska, A., Munteanu, V., Yang, H., Rotman, J., Tao, L., Balliu, B., Tseng, E., Eskin, E., Zhao, F., Mohammadi, P., P Łabaj, P., & Mangul, S. (2023). RNA-seq data science: From raw data to effective interpretation. *Frontiers in Genetics*, *14*, 997383. https://doi.org/10.3389/fgene.2023.997383

17. Diehl, A. D., Meehan, T. F., Bradford, Y. M., Brush, M. H., Dahdul, W. M., Dougall, D. S., He, Y., Osumi-Sutherland, D., Ruttenberg, A., Sarntivijai, S., Van Slyke, C. E., Vasilevsky, N. A., Haendel, M. A., Blake, J. A., & Mungall, C. J. (2016). The cell ontology 2016: Enhanced content, modularization, and ontology interoperability. *Journal of Biomedical Semantics*, *7*(1), 1–10. https://doi.org/10.1186/s13326-016-0088-7

18. Edgar, R., & Barrett, T. (2006). NCBI GEO standards and services for microarray data [1]. *Nature Biotechnology*, *24*(12), 1471–1472. https://doi.org/10.1038/nbt1206-1471

19. *FAANG*. (2025). https://data.faang.org/

20. Feijoó, M. P. P., Jardim, R., da Cruz, S. M. S., & Campos, M. L. M. (2020). Evaluating FAIRness of Genomic Databases. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12584 LNCS*(December), 128–137. https://doi.org/10.1007/978-3-030-65847-2_12

21. Feldman, E., Goutman, S., Petri, S., Mazzini, L., Savelieff, M., Shaw, P., & G, S. (2022). Amyotrophic lateral sclerosis. *Lancet*, *400*(10360), 1363–1380. https://doi.org/10.1016/S0140-6736(22)01272-7

22. Füllgrabe, A., George, N., Green, M., Nejad, P., Aronow, B., Fexova, S. K., Fischer, C., Freeberg, M. A., Huerta, L., Morrison, N., Scheuermann, R. H., Taylor, D., Vasilevsky, N., Marioni, J., Teichmann, S., Brazma, A., & Papatheodorou, I. (2020). Guidelines for reporting single-cell RNA-Seq experiments. *Nature Biotechnology*, *38*, 1384–1386.

23. Galbraith, D. W. (2006). The daunting process of MIAME. *Nature*, *444*(November), 31.

24. Glasziou, P., Altman, D. G., Bossuyt, P., Boutron, I., Clarke, M., Julious, S., Michie, S., Moher, D., & Wager, E. (2014). Reducing waste from incomplete or unusable reports of biomedical research. *The Lancet*, *383*(9913), 267–276. https://doi.org/10.1016/S0140-6736(13)62228-X

25. Gonçalves, R. S., & Musen, M. A. (2019). Analysis: The variable quality of metadata about biological samples used in biomedical experiments. *Scientific Data*, *6*(February), 1–15. https://doi.org/10.1038/sdata.2019.21

26. Gonçalves, R. S., O'Connor, M. J., Martínez-Romero, M., Graybeal, J., & Musen, M. A. (2017). Metadata in the BioSample online repository are impaired by numerous anomalies. *CEUR Workshop Proceedings*, *1931*, 39–46.

27. Grima, N., Liu, S., Southwood, D., Henden, L., Smith, A., Lee, A., Rowe, D. B., D'Silva, S., Blair, I. P., & Williams, K. L. (2023). RNA sequencing of peripheral blood in amyotrophic lateral sclerosis reveals distinct molecular subtypes: Considerations for biomarker discovery. *Neuropathology and Applied Neurobiology*, *49*(6), e12943. https://doi.org/10.1111/nan.12943

28. Gundersen, S., Boddu, S., Capella-Gutierrez, S., Drabløs, F., Fernández, J. M., Kompova, R., Taylor, K., Titov, D., & Zerbino, D. (2021). Recommendations for the FAIRification of genomic track metadata. *F1000Research*, *10*, 1–22. https://doi.org/10.12688/f1000research.28449.1

29. *HCA Data Portal*. (2025). https://data.humancellatlas.org/

30. Huang, Y.-N., Rajesh, A., Ayyala, R., Sarkar, A., Guo, R., Ling, E., Nakashidze, I., Wong, M. Y., Hu, J., Yu, D., Peng, Q., Scheg, G., Patel, K. K., Ramesh, T., Wang, K., Yadav, A., Liu, F., Mehta, J. H., Boldirev, G., … Mangul, S. (2022). The systematic assessment of completeness of public metadata accompanying omics studies. *BioRxiv*, 2021.11.22.469640. https://doi.org/10.1101/2021.11.22.469640

31. Juty, N., Wimalaratne, S. M., Soiland-Reyes, S., Kunze, J., Goble, C. A., & Clark, T. (2020). Unique, Persistent, Resolvable: Identifiers as the Foundation of FAIR. *Data Intelligence*, *2*(1–2), 30–39. https://doi.org/10.1162/dint_a_00025

32. Kiaei, L., & Kiaei, M. (2022). RNA as a Source of Biomarkers for Amyotrophic Lateral Sclerosis. *Metabolic Brain Disease*, *37*(3), 1697–1702. https://doi.org/10.1007/s11011-021-00738-z

33. Kim, R. B., Irvin, C. W., Tilva, K. R., & Mitchell, C. S. (2015). State of the field: An informatics-based systematic review of the SOD1-G93A amyotrophic lateral sclerosis transgenic mouse model. *Amyotrophic Lateral Sclerosis & Frontotemporal Degeneration*, *17*(1–2), 1–14. https://doi.org/10.3109/21678421.2015.1047455

34. Kumar, P., Halama, A., Hayat, S., Billing, A. M., Gupta, M., Yousri, N. A., Smith, G. M., & Suhre, K. (2015). MetaRNA-Seq: An interactive tool to browse and annotate metadata from RNA-Seq studies. *BioMed Research International*, *2015*, 1–6. https://doi.org/10.1155/2015/318064

35. Lee, H., Bartolome-Casado, R., Salas, A., Lance, C., & Kimler, K. (2025). *Integrated Gut Cell Atlas Metadata Submission*. https://hca_gut_cell_atlas.storage.googleapis.com/metadata_correctness.html

36. Leipzig, J., Nüst, D., Hoyt, C. T., Ram, K., & Greenberg, J. (2021). The role of metadata in reproducible computational research. *Patterns (New York, N.Y.)*, *2*(9), 100322. https://doi.org/10.1016/j.patter.2021.100322

37. Martínez-Romero, M., Jonquet, C., O'Connor, M. J., Graybeal, J., Pazos, A., & Musen, M. A. (2017). NCBO Ontology Recommender 2.0: An enhanced approach for biomedical ontology recommendation. *Journal of Biomedical Semantics*, *8*(1), 1–22. https://doi.org/10.1186/s13326-017-0128-y

38. Matsuda, T. (2021). Importance of experimental information (metadata) for archived sequence data: case of specific gene bias due to lag time between sample harvest and RNA protection in RNA sequencing. *PeerJ*, *9*, e11875. https://doi.org/10.7717/peerj.11875

39. Metadata Technology North America. (2025). *MetaSheet*. https://github.com/mtna/metasheet

40. Muneer, M. T. A., Rajan, M. S., Shenoy, A., Hegde, K., Koshy, S., & Jan-mar, V. (2014). A Comprehensive Review on Amyotrophic Lateral Sclerosis. *International Journal of Pharmaceutical and Chemical Sciences*, *3*(1), 19–22.

41. Rajesh, A., Chang, Y., Abedalthagafi, M. S., Wong-Beringer, A., Love, M. I., & Mangul, S. (2021). Improving the completeness of public metadata accompanying omics studies. *Genome Biology*, *22*(1), 1–5. https://doi.org/10.1186/s13059-021-02332-z

42. Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., Clevers, H., Deplancke, B., Dunham, I., Eberwine, J., Eils, R., Enard, W., Farmer, A., Fugger, L., Göttgens, B., … Yosef, N. (2017). The human cell atlas. *ELife*, *6*, 1–30. https://doi.org/10.7554/eLife.27041

43. Rocca-Serra, P., Gu, W., Ioannidis, V., Abbassi-Daloii, T., Capella-Gutierrez, S., Chandramouliswaran, I., Splendiani, A., Burdett, T., Giessmann, R. T., Henderson, D., Batista, D., Emam, I., Gadiya, Y., Giovanni, L., Willighagen, E., Evelo, C., Gray, A. J. G., Gribbon, P., Juty, N., … Sansone, S. A. (2023). The FAIR Cookbook - the essential resource for and by FAIR doers. *Scientific Data*, *10*(1), 1–12. https://doi.org/10.1038/s41597-023-02166-3

44. Rustici, G., Williams, E., Barzine, M., Brazma, A., Bumgarner, R., Chierici, M., Furlanello, C., Greger, L., Jurman, G., Miller, M., Ouellette, B. F. F., Quackenbush, J., Reich, M., Stoeckert, C. J.,

Taylor, R. C., Trutane, S. C., Weller, J., Wilhelm, B., & Winegarden, N. (2021). Transcriptomics data availability and reusability in the transition from microarray to next-generation sequencing. *BioRxiv*. https://doi.org/10.1101/2020.12.31.425022

45. Sabot, F. (2022). On the importance of metadata when sharing and opening data. In *BMC genomic data* (Vol. 23, Issue 1, p. 79). https://doi.org/10.1186/s12863-022-01095-1

46. Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J. P., Sun, L., Turner, S., & Karsch-Mizrachi, I. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database : The Journal of Biological Databases and Curation*, *2020*. https://doi.org/10.1093/database/baaa062

47. Schriml, L. M., Mitraka, E., Munro, J., Tauber, B., Schor, M., Nickle, L., Felix, V., Jeng, L., Bearer, C., Lichenstein, R., Bisordi, K., Campion, N., Hyman, B., Kurland, D., Oates, C. P., Kibbey, S., Sreekumar, P., Le, C., Giglio, M., & Greene, C. (2019). Human Disease Ontology 2018 update: Classification, content and workflow expansion. *Nucleic Acids Research*, *47*(D1), D955–D962. https://doi.org/10.1093/nar/gky1032

48. Seep, L., Grein, S., Splichalova, I., Ran, D., Mikhael, M., Hildebrand, S., Lauterbach, M., Hiller, K., Ribeiro, D. J. S., Sieckmann, K., Kardinal, R., Huang, H., Yu, J., Kallabis, S., Behrens, J., Till, A., Peeva, V., Strohmeyer, A., Bruder, J., … Hasenauer, J. (2024). From Planning Stage Towards FAIR Data: A Practical Metadatasheet For Biomedical Scientists. *Scientific Data*, *11*(1), 1–14. https://doi.org/10.1038/s41597-024-03349-2

49. Sheffield, N. C., LeRoy, N. J., & Khoroshevskyi, O. (2023). Challenges to sharing sample metadata in computational genomics. *Frontiers in Genetics*, *14*, 1154198. https://doi.org/10.3389/fgene.2023.1154198

50. Shillum, C., Petro, J. A., Demeranville, T., Wijnbergen, I., Hershberger, S., & Simpson. (2021). *From Vision to Value: ORCID's 2022–2025 Strategic Plan*.

51. Simoneau, J., Dumontier, S., Gosselin, R., & Scott, M. S. (2021). Current RNA-seq methodology reporting limits reproducibility. *Briefings in Bioinformatics*, *22*(1), 140–145. https://doi.org/10.1093/bib/bbz124

52. Smail, C., & Montgomery, S. B. (2024). RNA Sequencing in Disease Diagnosis. *Annual Review of Genomics and Human Genetics*, *25*(1), 353–367. https://doi.org/10.1146/annurev-genom-021623-121812

53. Spellman, P. T., Miller, M., Stewart, J., Troup, C., Sarkans, U., Chervitz, S., Bernhart, D., Sherlock, G., Ball, C., Lepage, M., Swiatek, M., Marks, W., Goncalves, J., Markel, S., Iordan, D., Shojatalab, M., Pizarro, A., White, J., Hubley, R., … Brazma, A. (2002). Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biology*, *3*(9), 1–9. https://doi.org/10.1186/gb-2002-3-9-research0046

54. Starr, J., Castro, E., Crosas, M., Dumontier, M., Downs, R. R., Duerr, R., Haak, L. L., Haendel, M., Herman, I., Hodson, S., Hourclé, J., Kratz, J. E., Lin, J., Nielsen, L. H., Nurnberger, A., Proell, S., Rauber, A., Sacchi, S., Smith, A., … Clark, T. (2015). Achieving human and machine accessibility of cited data in scholarly publications. *PeerJ. Computer Science*, *1*. https://doi.org/10.7717/peerj-cs.1

55. Swinnen, B., & Robberecht, W. (2014). The phenotypic variability of amyotrophic lateral sclerosis. *Nature Reviews. Neurology*, *10*(11), 661–670. https://doi.org/10.1038/nrneurol.2014.184

56. Toker, L., Feng, M., & Pavlidis, P. (2016). Whose sample is it anyway? Widespread misannotation of samples in transcriptomics studies. *F1000Research*, *5*, 1–15. https://doi.org/10.12688/F1000RESEARCH.9471.1

57. Ulrich, H., Kock-Schoppenhauer, A.-K., Deppenwiese, N., Gött, R., Kern, J., Lablans, M., Majeed, R. W., Stöhr, M. R., Stausberg, J., Varghese, J., Dugas, M., & Ingenerf, J. (2022). Understanding

the Nature of Metadata: Systematic Review. *Journal of Medical Internet Research*, *24*(1), e25440. https://doi.org/10.2196/25440

58. van Gils, B. (2023). Metadata. In *Data in Context* (pp. 159–160). Springer, Cham. https://doi.org/https://doi.org/10.1007/978-3-031-35539-4_16

59. van Rheenen, W., van der Spek, R. A. A., Bakker, M. K., van Vugt, J. J. F. A., Hop, P. J., Zwamborn, R. A. J., de Klein, N., Westra, H.-J., Bakker, O. B., Deelen, P., Shireby, G., Hannon, E., Moisse, M., Baird, D., Restuadi, R., Dolzhenko, E., Dekker, A. M., Gawor, K., Westeneng, H.-J., … Veldink, J. H. (2021). Common and rare variant association analyses in amyotrophic lateral sclerosis identify 15 risk loci with distinct genetic architectures and neuron-specific biology. *Nature Genetics*, *53*(12), 1636–1648. https://doi.org/10.1038/s41588-021-00973-1

60. Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 1–9. https://doi.org/10.1038/sdata.2016.18

61. Wiltgen, M., & Tilz, G. P. (2007). DNA microarray analysis: principles and clinical impact. *Hematology (Amsterdam, Netherlands)*, *12*(4), 271–287. https://doi.org/10.1080/10245330701283967

62. Witwer, K. W. (2013). Data submission and quality in microarray-based MicroRNA profiling. *Clinical Chemistry*, *59*(2), 392–400. https://doi.org/10.1373/clinchem.2012.193813

63. Wong, K.-C. (2019). Big data challenges in genome informatics. In *Biophysical reviews* (Vol. 11, Issue 1, pp. 51–54). https://doi.org/10.1007/s12551-018-0493-5

64. Zarei, S., Carr, K., Reiley, L., Diaz, K., Guerra, O., Altamirano, P. F., Pagani, W., Lodin, D., Orozco, G., & Chinea, A. (2015). A comprehensive review of amyotrophic lateral sclerosis. *Surgical Neurology International*, *6*, 171. https://doi.org/10.4103/2152-7806.169561

65. Ziff, O. J., Neeves, J., Mitchell, J., Tyzack, G., Martinez-Ruiz, C., Luisier, R., Chakrabarti, A. M., McGranahan, N., Litchfield, K., Boulton, S. J., Al-Chalabi, A., Kelly, G., Humphrey, J., & Patani, R. (2023). Integrated transcriptome landscape of ALS identifies genome instability linked to TDP-43 pathology. *Nature Communications*, *14*(1), 2176. https://doi.org/10.1038/s41467-023-37630-6

Appendices

Appendix 1

Structured Metadata Template for RNA-Seq Submission to GEO

Entity	Field	Description
Study	Title	Unique title (maximum of 255 characters) that describes the overall study. A working title from the associated manuscript may be suitable.
Study	Summary (Abstract)	Thorough description of the goals and objectives of this study. A working abstract from the associated manuscript may be suitable. Include as much text as necessary.
Study	Experimental design	Description of the type of SAMPLES included in the submission. For example: The experimental conditions and variables under investigation, if replicates are included, are there control and/or reference Samples, etc.
Study	Contributor	Firstname, Middlename initial, Lastname Or Firstname, Lastname Each contributor on a separate line. Add as many contributor lines as needed.
Study	Supplementary file	If you submit a processed data file corresponding to multiple Samples, include the processed file name here.
Samples	Library name	In each row, provide a unique name for the library. Do not include sensitive information. Use only plain ASCII text for the names. Each row represents a GEO Sample record (GSMxxxxxxxx). The unique name will be displayed in the "Description" field of the GSM.
Samples	Title	Unique title (maximum of 120 characters) that describes the Sample. We suggest the convention: [biomaterial] [condition(s)] [replicate number]
Samples	Library strategy	A Sequence Read Archive (SRA)-specific field that describes the sequencing technique for each library. Enter one of the strategies from the drop-down list (click arrowhead on cell immediately below).
Samples	Organism	Provide a valid scientific name at species level (or lower rank, such as subspecies). Example**:** Mus musculus The name must return a SINGLE entry in the Taxonomy database: https://www.ncbi.nlm.nih.gov/taxonomy/ Add another "organism" column(s), if needed.
Samples	Tissue	
Samples	Cell line	
Samples	Cell type	

Entity	Field	Description
Samples	Genotype	
Samples	Treatment	
Samples	Batch	
Samples	Molecule	Type of molecule that was extracted from the biological material. Include one of the molecules from the drop-down list
Samples	Single or paired-end	Include "single" or "paired-end"
Samples	Instrument model	Instrument model used to sequence the library.
Samples	Description	Additional information not provided in the other fields, or broad descriptions that cannot be easily dissected into other fields.
Samples	Processed data file	Exact name of the file containing the processed data.
Samples	Raw file	Exact name of the file containing the raw data.

Appendix 2

MIAME-Based Metadata Template for Microarray and RNA-seq Studies

Table 1

MIAME Metadata standard: Experimental Design

Field	Description	Answer Type	Link
Author (submitter)	First and Last Name	Text	
Email Address		E-mail	
Institution Name		Text	
Links(URL)	Link to related publication or data repository	Link	
Experiment Title	A brief, descriptive single-sentence title for the experiment.	Text	
Description	A free-text format description of the experiment or a link to an electronically available publication	Text	
Type of the Experiment	Such as normal-versus-diseased comparison, time course, dose response, and so on	Text	
Experimental Variables	Describes the parameters or conditions tested, such as time, dose, genetic variation, or response to a treatment or compound	Text	
Experimental Variable Values	Specific values for experimental factors (e.g., dose level, time of treatment).	Text	
Replicates	Usage and types of replicates.	Multiple Choice	https://www.nature.com/documents/Biological_and_technical_replicates_guidelines.pdf
Quality-control	Such as dealing with low-complexity sequence-induced	Text	https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8

Field	Description	Answer Type	Link
	nonspecific hybridization		
Experimental Relationships	Describes the relationship between the samples and the arrays used in each hybridization.	Text	

Table 2

MIAME Metadata standard: Array Design

Field	Description	Answer Type	Link
Array ID	A unique identifier for each array used.	Text	
Array	Is your array commercial or custom?	Multiple Choice	
Provider	The provider or manufacturer of the array.	Text	
Array Platform Type	Describes the platform type (e.g., glass slides, nylon membrane).	Multiple Choice	
Surface Type	Describes the surface type of the array (e.g., type of material used, such as glass, silicon).	Multiple Choice	
Array Content	Describes each element or spot on the array, such as the probes or genes	Text	
Array Element Identification	Describes the properties of each element, such as gene identifiers or sequences	Text	

Table 3

MIAME Metadata standard: Sample Information

Field	Description	Answer Type	Link
Sample ID	A unique identifier for each sample.	Text	
Organism	Species (e.g., Homo sapiens, Mus musculus). NCBI Taxonomy.	Text	https://www.ncbi.nlm.nih.gov/taxonomy
Tissue/Cell Type	Type of tissue or cells used in the study. Uberon or Cell Ontology.	Text	https://www.ebi.ac.uk/ols/ontologies/uberon
Qualifier	Defines a characteristic of the sample (e.g., strain, age, treatment).	Multiple Choice	

Field	Description	Answer Type	Link
Value	The value associated with the Qualifier.	Text	
Source	The source of the information (e.g., database, lab records).	Text	https://www.disease-ontology.org/
Experimental Factors	Key experimental variables (e.g., drug dose, time point).	Text	
Sample Treatment Type	Describes whether the treatment was applied in vitro, in vivo, or not applied.	Multiple Choice	
Sample Treatment	Describes the treatment applied to sample	Text	
Sample Preparation	Describes the method used to prepare the samples for analysis	Multiple Choice	
Sample Labeling	Describes the labeling method for RNA or DNA.	Text	

Table 4

MIAME Metadata standard: Hybridization

Field	Description	Answer Type	Link
Hybridization Protocol	A free-text description of the hybridization protocol	Text	
Hybridization Conditions	Parameters for hybridization (e.g., temperature, volume, wash procedure).	None	
Hybridization Solution	Salt and detergent concentrations	Text	
Blocking Agent	Agent used to block non-specific binding during hybridization.	Text	
Wash Procedure	Procedure to wash hybridized samples	Text	
Wash Temperature	Temperature of wash (°C)	Number	
Wash Time	Duration of the wash (minutes)	Number	
Wash Buffer	Composition of the wash buffer	Text	
Quantity of Labeled Target Used	Amount of labeled target used	Text	
Hybridization Time	Duration of hybridization (hours?)	Number	
Hybridization Volume	Total volume of hybridization (µL?)	Number	
Hybridization Temperature	Temperature of hybridization (°C)	Number	

Field	Description	Answer Type	Link
Hybridization Equipment	Equipment used for hybridization	Text	

Table 5

MIAME Metadata standard: Measurements

Field	Description	Answer Type	Link
Raw Data	Original scans (e.g. TIFF, CEL, or GPR)	Files	
Quantification Matrix	Quantification matrices (e.g. from image analysis)	Spreadsheet	
Image Analysis Software	Software used to analyze microarray images	Text	
Methodology	Underlying methodology (e.g., algorithms)	Text	
Quantifications	Definitions of quantifications used	Text	
Gene Expression Matrix	Final gene expression matrix after normalization	Text	
Numerical Calculations	All numerical calculation specifications	Text	

Table 6

MIAME Metadata standard: Normalization Controls

Field	Description	Answer Type	Link
Normalization Strategy	General approach for normalization	Multiple Choice	
Normalization Method	Method used for normalization	Multiple Choice	
Quality Control	Quality control algorithms and strategies	Text	
Control Elements	Details of control elements in the experiment	Text	
Hybridization Extract Preparation	Preparation of control probes/samples before hybridization	Text	

Appendix 3

FAANG-Based Metadata Template for RNA-seq Studies

Table 1

FAANG Metadata standard: Standard

Field	Description	Answer Type
Sample Description	A brief description of the sample including species name.	string
Material	The type of material being described.	ontology id
Project	State that the project is 'FAANG'.	constant
Secondary project	State the secondary project(s) that this data belongs to e.g. 'AQUA-FAANG', 'GENE-SWitCH' or 'BovReg'. Please use your official consortium shortened acronym if available. If your secondary project is not in the list, please contact the faang-dcc helpdesk to have it added. If your project uses the FAANG data portal project slices (https://data.faang.org/projects) then this field is required to ensure that your data appears in the data slice.	string
Availability	Either a link to a web page giving information on sample availability (who to contact and if the sample is available), or a e-mail address to contact about availability. E-mail addresses should be prefixed with 'mailto:', e.g. 'mailto:samples@example.ac.uk'. In either case, long term support of the web page or e-mail address is necessary. Group e-mail addresses are preferable to indiviudal.	uri
Same as	BioSample ID for an equivalent sample record, created before the FAANG metadata specification was available. This is optional and not intended for general use, please contact the data coordination centre (faang-dcc@ebi.ac.uk) before using it.	string

Table 2

FAANG Metadata standard: Organism

Field	Description	Answer Type
Organism	NCBI taxon ID of organism.	ontology id
Sex	Animal sex, described using any child term of PATO_0000047.	ontology id
Birth date	Birth date, in the format YYYY-MM-DD, or YYYY-MM where only the month is known. For embryo samples record 'not applicable'.	string
Breed	Animal breed, described using the FAANG breed description guidelines (http://bit.ly/FAANGbreed). Should be considered mandatory for terrestiral species, for aquatic species record 'not applicable'.	ontology id
Health status	Healthy animals should have the term normal, otherwise use the as many disease terms as necessary from EFO.	ontology id
Diet	Organism diet summary, more detailed information will be recorded in the associated protocols. Particuarly important for projects with controlled diet treatements. Free text field, but ensure standardisation within each study.	string
Birth location	Name of the birth location.	string
Birth location latitude	Latitude of the birth location in decimal degrees.	number
Birth location longitude	Longitude of the birth location in decimal degrees.	number
Birth weight	Birth weight, in kilograms or grams.	number
Placental weight	Placental weight, in kilograms or grams.	number
Pregnancy length	Pregnancy length of time, in days, weeks or months.	number
Delivery timing	Was pregnancy full-term. early or delayed.	string
Delivery ease	Did the delivery require assistance.	string
Pedigree	A link to pedigree information for the animal.	uri
Child of	Sample name or Biosample ID for sire/dam. Required if related animals are part of FAANG, e.g. quads.	string

Table 3

FAANG Metadata standard: Organoid

Field	Description	Answer Type
Organ model	Organ for which this organoid is a model system e.g. 'heart' or 'liver'. High level organ term.	ontology id
Organ part model	Organ part for which this organoid is a model system e.g. 'bone marrow' or 'islet of Langerhans'. More specific part of organ.	ontology id
Freezing date	Date that the organoid was frozen.	string
Freezing method	Method of freezing of organoid. Temperatures are in celsius. 'Frozen, vapor phase' refers to storing samples above liquid nitrogen in the vapor.	string
Freezing protocol	A link to the protocol for freezing.	uri
Number of frozen cells	Number of organoids cells that were frozen.	number
Organoid passage	Number of passages. Passage 0 is the plating of cells to create the organoid	number
Organoid passage protocol	A link to the protocol for organoid passage.	uri
Organoid culture and passage protocol	Protocol for the culture and passage of organoids, growth environment (matrigel or other); incubation temperature and oxygen level are expected in this protocol	uri
Growth environment	Growth environment in which the organoid is grown. e.g. 'matrigel', 'liquid suspension' or 'adherent'.	string
Type of organoid culture	Whether the organoid culture two dimensional or three dimensional.	string
Organoid morphology	General description of the organoid morphology. e.g. 'Epithelial monolayer with budding crypt-like domains' or 'Optic cup structure'. Be consistent within your project if multiple similar samples.	string
Derived from	Sample name or BioSample ID for a specimen or organoid record.	string

Table 4

FAANG Metadata standard: Experiment Standard

Field	Description	Answer Type
Secondary project	State the secondary project(s) that this data belongs to e.g. 'AQUA-FAANG', 'GENE-SWitCH' or 'BovReg'. Please use your official consortium shortened acronym if available. If your secondary project is not in the list, please contact the faang-dcc helpdesk to have it added. If your project uses the FAANG data portal project slices (https://data.faang.org/projects) then this field is required to ensure that your data appears in the data slice.	string
Assay type	The type of experiment performed.	string
Sample storage	How the sample was stored. Temperatures are in celsius. 'Frozen, vapor phase' refers to storing samples above liquid nitrogen in the vapor.	string
Sample storage processing	How the sample was prepared for storage.	string
Sampling to preparation interval	How long between the sample being taken and the assay experiment preperations commencing. If sample preperations were then left in intermediate stages after preperation commenced, for example as sheared chromatin, then this should be made clear in your protocols.	number
Experimental protocol	Link to the description of the experiment protocol, an overview of the full experiment, that can refer to the order in which other protocols were performed and any intermediate steps not covered by the extraction or assay specific protocols.	uri
Extraction protocol	Link to the protocol used to isolate the extract material.	uri
Library preparation location	Location where library preperation was performed.	string
Library preparation location longitude	Longitude of the library prep location in decimal degrees.	number
Library preparation location latitude	Latitude of the library prep location in decimal degrees.	number
Library preparation date	Date on which the library was prepared.	string
Sequencing location	Location of where the sequencing was performed.	string

Field	Description	Answer Type
Sequencing location longitude	Longitude of the sequencing location in decimal degrees.	number
Sequencing location latitude	Latitude of the sequencing location in decimal degrees.	number
Sequencing date	Date sequencing was performed.	string

Table 5

FAANG Metadata standard: RNA Sequencing

Field	Description	Answer Type
Experiment target	What the experiment was trying to find, list the text rather than ontology link, for example 'polyA RNA'.	ontology id
Rna preparation 3' adapter ligation protocol	Link to the protocol for 3’ adapter ligation used in preparation.	uri
Rna preparation 5' adapter ligation protocol	Link to the protocol for 5’ adapter ligation used in preparation.	uri
Library generation pcr product isolation protocol	Link to the protocol for isolating pcr products used for library generation.	uri
Preparation reverse transcription protocol	Link to the protocol for reverse transcription used in preparation.	uri
Library generation protocol	Link to the protocol used to generate the library.	uri
Read strand	For strand specific protocol, specify which mate pair maps to the transcribed strand or Report 'non-stranded' if the protocol is not strand specific. For single-ended sequencing: use 'sense' if the reads should be on the same strand as the transcript, 'antisense' if on opposite strand. For paired-end sequencing: 'mate 1 sense' if mate 1 should be on the same strand as the transcript, 'mate 2 sense' if mate 2 should be on the same strand as the transcript.	string

Appendix 4

HCA-Based Metadata Template for RNA-seq Studies (*HCA Data Portal*, 2025)

Field	Description	Answer Type	Link
Title	A brief, descriptive single-sentence title for the experiment.	Text	
Description	A free-text format description of the experiment or a link to an electronically available publication	Text	
Author (submitter)	First and Last Name	Text	
Email Address		E-mail	
Institution Name		Text	
Links(URL)	Publication/preprint DOI: The publication digital object identifier (doi) for the protocol. If no pre-print nor publication exists, please write 'not applicable'.	Link	
Consortia			
title	This text describes and differentiates the dataset from other datasets in the same collection. It is strongly recommended that each dataset title in a collection is unique and does not depend on other metadata such as a different assay to disambiguate it from other datasets in the collection.	Text	
study_pi	Principal Investigator(s) leading the study where the data is/was used	Text	
batch_condition	Note: Name of the covariate that confers the dominant batch effect in the data as judged by the data contributor. The name provided here should be the label by which this covariate is stored in the AnnData object.	Text	
default_embedding	The value must match a key to an embedding in obsm for the embedding to display by default in CELLxGENE Explorer.	Text	
comments	Other technical or experimental covariates that could affect the quality or batch of the sample. Must not contain identifiers.	Text	

Field	Description	Answer Type	Link
	This field is designed to capture potential challenges for data integration not captured elsewhere.		
raw counts			
normalized counts			
protocol_url	The protocols.io URL (if none exists, please use the BioRxiv URL) for the full experimental protocol; or if multiple protocols exist please list them e.g. sample preparation protocol / sequencing protocol.	Text	
donor_id	This must be free-text that identifies a unique individual that data were derived from	Text	
sample_id	Identification number of the sample. This is the fundamental unit of sampling the tissue (the specimen taken from the subject), which can be the same as the 'subject_ID', but is often different if multiple samples are taken from the same subject. Note: this is NOT a unit of multiplexing of donor samples, which should be stored in "library".	Text	
institute	Institution where the samples were processed.	Text	
sample_collection_site	The pseudonymised name of the site where the sample was collected.	Text	
sample_collection_relative_time_point	Time point when the sample was collected. This field is only needed if multiple samples from the same subject are available and collected at different time points. Sample collection dates (e.g. 23/09/22) cannot be used due to patient data protection, only relative time points should be used here (e.g. day3).	Text	
library_id	The unique ID that is used to track libraries in the investigator's institution (should align with the publication)	Text	

Field	Description	Answer Type	Link
library_id_repository	The unique ID used to track libraries from one of the following public data repositories: EGAX*, GSM*, SRX*, ERX*, DRX, HRX, CRX	Text	
author_batch_notes	Encoding of author knowledge on any further information related to likely batch effects.	Text	
organism_ontology_term_id	The name given to the type of organism, collected in NCBITaxon:0000 format.	Text	https://www.ncbi.nlm.nih.gov/taxonomy
manner_of_death	Manner of death classification based on the Hardy Scale or 'unknown' or 'not applicable': ● Category 1 = Violent and fast death Deaths due to accident, blunt force trauma or suicide, terminal phase estimated at < 10 min. ● Category 2 = Fast death of natural causes -Sudden unexpected deaths of people who had been reasonably healthy, after a terminal phase estimated at < 1 hr (with sudden death from a myocardial infarction as a model cause of death for this category) ● Category 3 = Intermediate death - Death after a terminal phase of 1 to 24 hrs (not classifiable as 2 or 4); patients who were ill but death was unexpected ● Category 4 = Slow death - Death after a long illness, with a terminal phase longer than 1 day (commonly cancer or chronic pulmonary disease); deaths that are not unexpected ● Category 0 =Ventilator Case - All cases on a ventilator immediately before death ● Unknown = The cause of death is unknown ● Not applicable = Subject is alive [Please leave this field as blank for embryonic/fetal tissue]	Multiple Choice	
sample_source	The study subgroup that the participant belongs to. This indicates whether the participant was a surgical donor (this includes patients providing blood samples or	Multiple Choice	

Field	Description	Answer Type	Link
	biopsies), a postmortem donor, or an organ donor.		
sex_ontology_term_id	Reported sex of the donor.	Multiple Choice	
sample_collection_method	The method the sample was physically obtained from the donor.	Multiple Choice	
tissue_type	Whether the tissue is "tissue", "organoid", or "cell culture".	Multiple Choice	
sampled_site_condition	Whether the site is considered healthy, diseased or adjacent to disease.	Multiple Choice	
tissue_ontology_term_id	The detailed anatomical location of the sample, please provide a specific UBERON term.	Text	https://www.ebi.ac.uk/ols/ontologies/uberon https://github.com/obophenotype/uberon
tissue_free_text	The detailed anatomical location of the sample - this does not have to tie to an ontology term	Text	
sample_preservation_method	Indicating if tissue was frozen, or not, at any point before library preparation	Text	
suspension_type	Specifies whether the sample contains single cells or single nuclei data.This must be "cell", "nucleus", or "na". This must be the correct type for the corresponding assay: ● 10x transcription profiling [EFO:0030080] and its children = "cell" or "nucleus" ● ATAC-seq [EFO:0007045] and its children = "nucleus" ● BD Rhapsody Whole Transcriptome Analysis [EFO:0700003] = "cell" ● BD Rhapsody Targeted mRNA [EFO:0700004] = "cell" ● CEL-seq2 [EFO:0010010] = "cell" or "nucleus" ● CITE-seq [EFO:0009294] and its children = "cell" ● DroNc-seq [EFO:0008720] = "nucleus" ● Drop-seq [EFO:0008722] = "cell" or "nucleus" ● GEXSCOPE technology [EFO:0700011] = "cell" or "nucleus" ● inDrop [EFO:0008780]	Multiple Choice	
cell_enrichment	Specifies the cell types targeted for enrichment or depletion beyond the selection of live		http://www.ebi.ac.uk/ols4/ontologies/cl

Field	Description	Answer Type	Link
	cells. This must be a Cell Ontology (CL) term. For cells that are enriched, list the CL code followed by a "+". For cells that were depleted, list the CL code followed by a "-". If no enrichment or depletion occurred, please use 'na' (not applicable)		
cell_viability_percentage	If measured, per sample cell viability before library preparation (as a percentage).	Number	
cell_number_loaded	Estimated number of cells loaded for library construction.	Number	
sample_collection_year	Year of sample collection. Should not be detailed further(to exact month and day), to prevent identifiability	Number	
assay_ontology_term_id	Platform used for single cell library construction. This must be an EFO term and either: ● "EFO:0002772" for assay by molecule or preferably its most accurate child ● "EFO:0010183" for single cell library construction or preferably its most accurate child ● An assay based on 10X Genomics products should either be "EFO:0008995" for 10x technology or preferably its most accurate child. An assay based on SMART (Switching Mechanism at the 5' end of the RNA Template) or SMARTer technology SHOULD either be "EFO:0010184" for Smart-like or preferably its most accurate child. Recommended: 10x 3' v2 "EFO:0009899" 10x 3' v3 "EFO:0009922" 10x 5' v1 "EFO:0011025" 10x 5' v2 "EFO:0009900" Smart-seq2 "EFO:0008931" Visium Spatial Gene Expression "EFO:0010961"	Text	https://www.ebi.ac.uk/ols4/ontologies/efo
library_preparation_batch	Indicating which samples' libraries were prepared in the same chip/plate/etc., e.g. batch1, batch2	Text	

Field	Description	Answer Type	Link
library_sequencing_run	The identifier (or accession number) that indicates which samples' libraries were sequenced in the same run.	Text	
sequenced_fragment	Which part of the RNA transcript was targeted for sequencing.	Multiple Choice	
sequencing_platform	Platform used for sequencing. EFO	Text	https://www.ebi.ac.uk/ols/ontologies/efo/terms?iri=http%3A%2F%2Fwww.ebi.ac.uk%2Fefo%2FEFO_0002699
is_primary_data	This must be True if this is the canonical instance of this cellular observation and False if not. This is commonly False for meta-analyses reusing data or for secondary views of data.	Multiple Choice	
reference_genome	Reference genome used for alignment	Multiple Choice	
gene_annotation_version	Ensembl release version accession number. Some common codes include: GRCh38.p12 = GCF_000001405.38 GRCh38.p13 = GCF_000001405.39 GRCh38.p14 = GCF_000001405.40	Text	http://www.ensembl.org/info/website/archives/index.html) or NCBI/RefSeq
alignment_software	Protocol used for alignment analysis, please specify which version was used e.g. cell ranger 2.0, 2.1.1 etc.	Text	
intron_inclusion	Were introns included during read counting in the alignment process?	Multiple Choice	
author_cell_type	Encoding of author intuition of cellular annotation in the dataset	Text	
cell_type_ontology_term	Cell Ontology (CL) term. This must be a Cell Ontology (CL) term (http://www.ebi.ac.uk/ols4/ontologies/cl). If no appropriate high-level term can be found or the cell type is unknown, then it is strongly recommended to use 'unknown'. The following terms must not be used: "CL:0000255" for eukaryotic cell; "CL:0000257" for Eumycetozoan cell; "CL:0000548" for animal cell	Text	http://www.ebi.ac.uk/ols4/ontologies/cl