Bioinformatics Advances, 2026, **6**(1), vbag092
https://doi.org/10.1093/bioadv/vbag092
Published: 28 March 2026
Application Note | Software

OXFORD

DoReMiTra: an R/Bioconductor data package for orchestrating the analysis of radiation transcriptomic studies

Ahmed Salah[1,2], Sebastian Zahnreich[1], Federico Marini[2,3,*]

[1]Department of Radiation Oncology and Radiation Therapy, University Medical Center of the Johannes Gutenberg University Mainz, Mainz, 55131, Germany
[2]Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), University Medical Center of the Johannes Gutenberg University Mainz, Mainz, 55131, Germany
[3]Research Center for Immunotherapy (FZI), University Medical Center of the Johannes Gutenberg University Mainz, Mainz, 55131, Germany
*Corresponding author. Federico Marini, Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), University Medical Center of the Johannes Gutenberg University Mainz, Mainz, 55131, Germany; Research Center for Immunotherapy (FZI), University Medical Center of the Johannes Gutenberg University Mainz, Mainz, 55131, Germany. E-mail: marinif@uni-mainz.de
Associate Editor: Nurcan Tunçbağ

Abstract

Summary: **Understanding the molecular impact of ionizing radiation exposure is essential for both biomedical research and public health. Among the possible approaches to study this phenomenon, gene expression profiling via transcriptomics assays has been a valuable approach over the last decades to unravel the mechanisms of cellular responses to radiation. To our knowledge, there is no data package gathering well-curated radiation transcriptomic datasets covering microarrays and, more recently, RNA sequencing. Therefore, we present DoReMiTra, an R/Bioconductor data package that represents the first unified radiation transcriptomics dataset collection integrated with Bioconductor's ExperimentHub for efficient distribution. DoReMiTra standardizes and harmonizes sample-level metadata and provides pre-processed SummarizedExperiment objects to facilitate comparative analyses. Additionally, we introduce a lightweight Shiny app interface for interactive visualization and preliminary exploration. DoReMiTra serves as a valuable resource and tool in radiation research for benchmarking, integrative analyses, and biomarker discovery.**

Availability and implementation: **DoReMiTra is available under the MIT license at https://bioconductor.org/packages/DoReMiTra.**

Introduction

Ionizing radiation (IR) is a long-standing cornerstone of medical practice, essential for diagnostic imaging and the treatment of both benign and malignant diseases. However, despite its significant clinical value, planned and especially accidental IR exposures still pose health risks (Ron 2002). Notably, radiation biology research increasingly relies on transcriptomic profiling to uncover the molecular mechanisms underlying cellular responses to IR (Kabacik *et al.* 2011, Manning *et al.* 2017). However, the diversity of radiation qualities, each with distinct energy deposition patterns and biological effects, can lead to diverse transcriptomic responses depending on dose, dose-rate, radiation type, and biological context (Ismail *et al.* 2012, Michalettou *et al.* 2021, Talapko *et al.* 2024). Therefore, gene expression profiling has emerged as a rapid and sensitive tool for dose estimation and toxicity prediction in clinical, biodosimetry, and triage applications (Dressman *et al.* 2007, Meadows *et al.* 2008).

Numerous gene expression studies have investigated these effects across human and animal models (El-Saghire *et al.* 2013, Hall *et al.* 2017). Still, data reuse is hindered by inconsistent metadata annotation, variable preprocessing methods, and a lack of accessibility, which makes it challenging to define experimental factors for statistical analysis (Koesten *et al.*, 2020, Bustos *et al.* 2026). To date, no resource integrates radiation transcriptomics datasets into a unified structure. Although some existing resources provide radiation-associated gene signatures, they do not distribute fully harmonized transcriptomic datasets suitable for direct computational analysis (Manem and Dhawan 2019, Sagkrioti *et al.* 2022).

To address this gap, we present DoReMiTra (radiation DOse REsponse Measured In TRAnscriptomics), a comprehensive R/Bioconductor data package that aggregates blood transcriptomic datasets from experiments involving photons (X-rays, γ-rays) and neutrons, providing access to well-documented and preprocessed transcriptomic datasets. We focused mainly on datasets generated from exposed blood samples since it is a very radiosensitive tissue that is always exposed to radiation in both medical and accidental exposure scenarios (Rothkamm *et al.* 2013, Heylmann *et al.* 2014). The package includes

Received: 7 January 2026. **Revised:** 7 March 2026. **Accepted:** 13 March 2026

datasets primarily retrieved from GEO via GEOquery (Davis and Meltzer 2007), unified under a common structure, with metadata covering radiation type, dose, time point, sex, organism, and experiment setting (e.g., in vivo, ex vivo). This data package not only streamlines access to a curated radiation transcriptomic landscape within ExperimentHub, but also provides all included datasets as SummarizedExperiment (SE) objects to allow seamless integration with R/Bioconductor workflows (Robinson *et al.* 2010, Love *et al.* 2014, Ritchie *et al.* 2015, Ludt *et al.* 2022), simplifying the comparability across studies and supporting reproducible analysis pipelines. This integration would make it feasible for biologists, bioinformaticians, and statisticians to conduct the analysis and integration without the hassle of metadata inconsistency across studies. DoReMiTra, accompanied by a Shiny app for exploratory visualization, serves as a valuable resource in radiation research and a tool for developing biomarkers of radiation exposure in the context of radiation therapy, radiation oncology, and biodosimetry, and encourages reproducibility and data reuse in radiation transcriptomic research (Wilkinson *et al.*, 2016, Fraga-González *et al.* 2025).

Data collection and processing

Many publicly available radiation transcriptomic datasets contain non-standardized or inconsistent metadata. In some studies, even samples from the same experiment are annotated with heterogeneous terminology or are described by overly long, unstructured text strings. These inconsistencies hinder the reliable extraction of key experimental factors such as dose, time point, or sex, and they complicate the construction of appropriate statistical design formulas for downstream analyses in packages such as limma, DESeq2, or edgeR.

To address this, the first building block of our package build was collecting datasets from irradiated blood samples of human or mouse, primarily from GEO, based on relevance to biodosimetry. Priority was given to experiments involving X-rays, γ-rays, and neutrons, with well-defined dose and time point information. The GEO queries to obtain all the datasets are available on the following link: https://github.com/AhmedSAHassan/DoReMiTra/blob/devel/inst/extdata/Geo_Queries.qmd.

All the information about each dataset, including accession number, author name, type of radiation, dose, experimental setting, etc., was collected in a tabular format. The datasets were then programmatically fetched by their accession number and processed into SE objects using standardized pipelines. Sample-level and feature-level metadata, reported in the respective colData and rowData slots, were manually curated with particular focus on adopting unified terminologies across all the datasets, including Platform, Radiation_type, Dose, Organism, Time_point, Sex, and Exp_setting, to facilitate the statistical analysis and integration. When metadata fields were missing or inconsistently reported, information was cross-validated using associated publications.

The RNA-seq dataset is provided as raw counts, whereas the normalization methods for the microarray datasets vary across studies, according to the decisions made by the original investigators due to platform-specific differences. Accordingly, a link to each dataset on GEO is included in the metadata of the corresponding SE object.

The current release includes 36 datasets, providing a high number of samples for integrative analysis. For example, the package provides 14 datasets for ex vivo γ-irradiated human blood samples with a total of 753 samples included. More datasets will be regularly included in the collection as they are generated and made publicly available. The current collection of datasets covers different organisms (Homo sapiens, Mus musculus, Macaca mulatta), radiation types (X-ray, gamma ray, neutron), experimental settings (in vivo, ex vivo), time points post-exposure (minutes to days), sex (male and female), and dose (Gy).

A detailed processing pipeline can be found in the Rmarkdown file `make-data-DoReMiTra.Rmd` from the installed `scripts` folder of the DoReMiTra package.

The DoReMiTra workflow

The DoReMiTra functionality includes a series of steps. The package components and workflow are shown in Figure 1. The primary entry point for exploring the package is `list_DoReMiTra_datasets()`, which returns a metadata data frame summarizing all datasets available in the DoReMiTra collection, including key attributes such as title, organism, radiation type, experimental setting, and accession numbers. Secondly, `get_all_DoReMiTra_Datasets()` retrieves all datasets from ExperimentHub and returns them as a named list of SummarizedExperiment (SE) objects. This function is particularly useful for batch processing or for global exploration of all curated radiation transcriptomic datasets. File retrieval is handled through BiocFileCache, which enables efficient caching of files.

To facilitate intuitive dataset discovery, we implemented the function `list_DoReMiTra_metadata_fields()`, which explicitly enumerates all available metadata fields, including radiation type, experimental settings, organism, and tissue type across the compendium, together with their corresponding values. Users can refine their selection using `search_DoReMiTra_datasets()`, which filters available datasets based on metadata fields such as radiation type, organism, author name, or experimental setting, allowing rapid identification of datasets relevant to a specific research question.

Once a dataset of interest is selected, `get_DoReMiTra_data()` fetches the corresponding SE object from ExperimentHub within seconds. The function includes a `gene_symbol` argument, which, when set to TRUE, sets rownames to gene symbols when available. In cases of duplicated symbols, the gene symbol and probe ID are appended; if the gene symbol is missing (NA), the probe ID is used instead. Setting `gene_symbol` to FALSE retains probe IDs as rownames. Once the SE object is retrieved by this function, it can be directly explored via the companion DoReMiTraExplorer package, which provides a Shiny application for visualization and exploratory analysis of the DoReMiTra datasets. The `DoReMiTra_explorer()` function enables users to intuitively explore transcriptomic responses to radiation through a suite of customizable plots, including principal component analysis (PCA) with user-defined gene subsets and color schemes based on experimental factors, boxplots for expression distributions, dose–response gene plots, and heatmaps of highly variable genes. Users can adjust parameters such as the

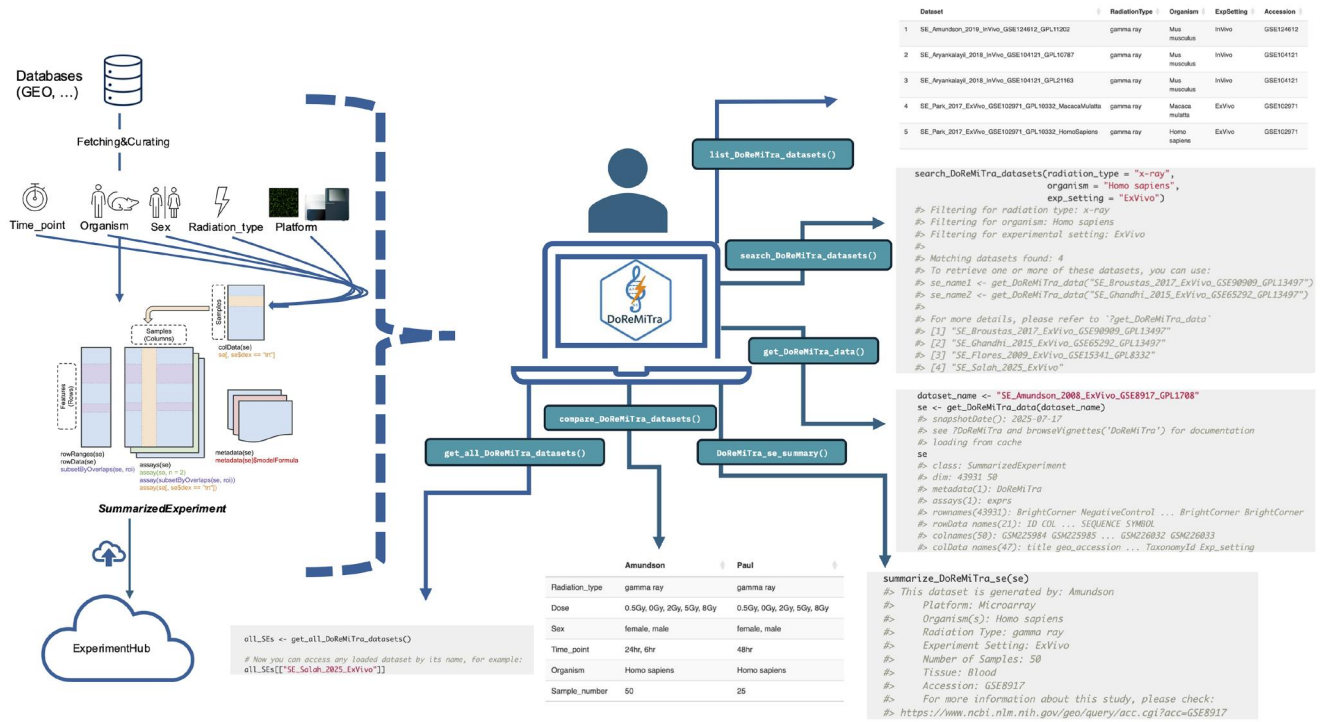

Figure 1 Graphical representation of the functionality and typical usage workflow for `DoReMiTra`. `get_DoReMiTra_data()` retrieves a dataset by name. `list_DoReMiTra_datasets()` views metadata for all available datasets. `search_DoReMiTra_datasets()` filters datasets by specified metadata fields. `compare_DoReMiTra_datasets()` compares metadata fields across datasets. `DoReMiTra_se_summary()` displays a summary of key metadata and GEO links. `get_all_DoReMiTra_datasets()` retrieves all datasets from the DoReMiTra collection.

number of genes displayed and the inclusion of clustering, offering flexibility for both overview and detailed inspection of radiation-induced transcriptional patterns. The Shiny app provided as a separate package can be found in the GitHub repository (https://github.com/AhmedSAHassan/DoReMiTraExplorer).

A quick overview of any retrieved dataset can be obtained using `summarize_DoReMiTra_se()`, which reports the number of samples, available metadata fields, and key experimental characteristics such as radiation type, organism, tissue, and platform. Finally, `compare_DoReMiTra_datasets()` allows users to compare two or more SE objects side-by-side by summarizing essential metadata attributes, including radiation type, dose, time point, and experimental settings.

An exemplary integrative analysis of datasets retrieved via DoReMiTra and visualized via DoReMiTraExplorer Shiny app is presented in Supplementary File 1, with the source code for the notebook provided in the GitHub repository https://github.com/AhmedSAHassan/DoReMiTraSup.

Discussion and future directions

DoReMiTra provides a one-stop shop of harmonized and accessible datasets for blood transcriptomic studies investigating the biological effects of various radiation types, including X-rays, γ-rays, and neutrons. By curating publicly available GEO datasets into standardized SE objects and embedding them within Bioconductor's ExperimentHub, DoReMiTra enables seamless access, reproducibility, and interoperability with the broader Bioconductor ecosystem. While some radiation transcriptomic resources, including RadBioBase (Sagkrioti *et al.* 2022), Rad-BioApp (Barker *et al.* 2021), and RadiationGeneSigDB (Manem and Dhawan 2019) have previously compiled radiation-related transcriptomic studies and represent valuable contributions to the radiation biology community, none of these provide access to the underlying expression data. DoReMiTra, together with the complementary DoReMiTraExplorer visualization app, bridges a critical gap in radiation biology by offering full programmatic access to the original expression data, a standardized structure, and harmonized metadata; it lays the groundwork for integrative analyses, reproducible benchmarking, and the identification of robust radiation-responsive molecular signatures.

Although DoReMiTra does not embed a full framework for statistical analysis of the provided datasets, it substantially lowers the barrier to performing radiation transcriptomics research by providing the largest curated collection of harmonized datasets in an analysis-ready format. Embedding predefined analysis functions could restrict methodological flexibility and limit compatibility with evolving best practices in transcriptomics. By providing well-annotated datasets as structured SE objects, users can seamlessly apply a wide range of established Bioconductor tools for differential expression analysis, pathway enrichment, visualization, and cross-study integration according to their specific research needs. We refer the reader to the analysis of Salah *et al.* (2025), described in detail in the repository (https://github.com/AhmedSAHassan/Phybion_ShortReads_Analysis), which illustrates how proper design specification is essential to extract robust radiation signatures, especially when working with human data.

DoReMiTra is distributed as two separate packages, a lightweight data package for curated dataset access, and a separate

package containing the Shiny app for interactive exploration to maintain stability, reducing the dependency burden if users only need programmatic access to the curated datasets. Future developments of DoReMiTra will include modules for cross-platform normalization, which would be particularly valuable for meta-analysis and machine learning workflows that require comparable expression scales across datasets. Similarly, batch correction is intentionally left to downstream analyses, allowing users to apply appropriate methods (e.g., ComBat, RUVSeq, sva) (Leek *et al.* 2012, Risso *et al.* 2014, Zhang *et al.* 2020). This modular approach maintains flexibility, enabling researchers to tailor normalization and integration strategies to their analytical needs while ensuring reproducibility and methodological transparency. Additionally, we plan to include dataset-specific gene signatures, which will further facilitate cross-study comparisons and the identification of conserved radiation-responsive transcriptional patterns across different exposure types and experimental systems. As the current version includes datasets from irradiated blood samples due to their biological relevance, future updates will include other tissue types to broaden the spectrum to answer different research questions.

Beyond facilitating immediate reuse of existing data, DoReMiTra encourages the community to adopt transparent and interoperable data practices, ultimately strengthening research in radiobiology, biodosimetry, and radiotherapy. Additionally, DoReMiTra can provide valuable resources for training and education due to its diverse collection of well-structured datasets for multiple comparisons. The package promotes FAIR data principles by being findable (via ExperimentHub search), accessible (with programmatic retrieval), interoperable (adopting the Bioconductor SE structure), and reusable (with curated metadata and documentation). Taken together, DoReMiTra offers a solution for accessing, exploring, and analyzing curated radiation transcriptomic datasets, enabling researchers to perform integrative studies and accelerate the discovery of radiation-responsive biomarkers.

Author contributions

A.S. Data curation, methodology, software, validation, writing—original draft, writing—review and editing S.Z. Supervision, funding acquisition, project administration, resources, writing—review and editing F.M. Conceptualization, methodology, software, supervision, funding acquisition, project administration, resources, writing—original draft, writing—review and editing

Supplementary material

Supplementary material is available at *Bioinformatics Advances* online.

Conflicts of interest

None declared.

Funding

This study was supported by the German Federal Ministry of Education and Research, Grant 02NUK084A (A.S.). The work of FM is also supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Projektnummer 318346496 - SFB1292/2 TP19N.

Data availability

The DoReMiTra package is available on Bioconductor under the MIT license at https://bioconductor.org/packages/DoReMiTra, including documentation and a vignette to showcase its usage; the companion web application to explore the included datasets is available at https://github.com/AhmedSAHassan/DoReMiTraExplorer. The source code for the analysis notebook presented in the Supplementary File is available in the GitHub repository https://github.com/AhmedSAHassan/DoReMiTraSup.

References

Barker R, Costes SV, Miller J *et al.* Rad-bio-app: a discovery environment for biologists to explore spaceflight-related radiation exposures. *NPJ Microgravity* 2021;**7**:15. https://doi.org/10.1038/s41526-021-00143-x

Bustos ML, Song K, Brochu HN *et al.* Impact of non-standardized reporting on reproducibility, usability, and integration in nasopharyngeal metagenomic research: a systematic review. *Front Microbiol* 2026;**17**:1707004. https://doi.org/10.3389/fmicb.2026.1707004

Davis S, Meltzer PS. Geoquery: a bridge between the gene expression omnibus (geo) and bioconductor. *Bioinformatics* 2007;**23**:1846–7. https://doi.org/10.1093/bioinformatics/btm254

Dressman HK, Muramoto GG, Chao NJ *et al.* Gene expression signatures that predict radiation exposure in mice and humans. *PLoS Med* 2007;**4**:e106. https://doi.org/10.1371/journal.pmed.0040106

El-Saghire H, Thierens H, Monsieurs P *et al.* Gene set enrichment analysis highlights different gene expression profiles in whole blood samples x-irradiated with low and high doses. *Int J Radiat Biol* 2013;**89**:628–38. https://doi.org/10.3109/09553002.2013.782448

Fraga-González G, van de Wiel H, Garassino F *et al.* Affording reusable data: recommendations for researchers from a data-intensive project. *Sci Data* 2025;**12**:258. https://doi.org/10.1038/s41597-025-04565-0

Hall J, Jeggo PA, West C *et al.* Ionizing radiation biomarkers in epidemiological studies – an update. *Mutat Res/Rev Mutat Res* 2017;**771**:59–84. https://doi.org/10.1016/j.mrrev.2017.01.001

Heylmann D, Rödel F, Kindler T *et al.* Radiation sensitivity of human and murine peripheral blood lymphocytes, stem and progenitor cells. *Biochim Biophys Acta* 2014;**1846**:121–9. https://doi.org/10.1016/j.bbcan.2014.04.009

Ismail AH, Hamad MA, Harki EMT. Radiation exposure of leukemia blood samples and its impacts on the density of RBC, WBC, and PLT: in vitro. *OJBIPHY* 2012;**02**:130–6. https://doi.org/10.4236/ojbiphy.2012.24016

Kabacik S, Mackay A, Tamber N *et al.* Gene expression following ionising radiation: identification of biomarkers for dose estimation and prediction of individual response. *Int J Radiat Biol* 2011;**87**:115–29. https://doi.org/10.3109/09553002.2010.519424

Koesten L, Vougiouklis P, Simperl E *et al.* Dataset reuse: toward translating principles to practice. *Patterns* 2020;**1**:100136. https://doi.org/10.1016/j.patter.2020.100136

Leek JT, Johnson WE, Parker HS *et al.* The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 2012;**28**:882–3. https://doi.org/10.1093/bioinformatics/bts034

Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550. https://doi.org/10.1186/s13059-014-0550-8

Ludt A, Ustjanzew A, Binder H *et al.* Interactive and reproducible workflows for exploring and modeling RNA-seq data with pcaexplorer, ideal, and genetonic. *Curr Protoc* 2022;**2**:e411. https://doi.org/10.1002/cpz1.411

Manem VS, Dhawan A. Radiationgenesigdb: a database of oxic and hypoxic radiation response gene signatures and their utility in pre-clinical research. *Br J Radiol* 2019;**92**:20190198. https://doi.org/10.1259/bjr.20190198

Manning G, Tichý A, Sirák I *et al.* Radiotherapy-associated long-term modification of expression of the inflammatory biomarker genes ARG1, BCL2L1, and MYC. *Front Immunol* 2017;**8**:412. https://doi.org/10.3389/fimmu.2017.00412

Meadows SK, Dressman HK, Muramoto GG *et al.* Gene expression signatures of radiation response are specific, durable and accurate in mice and humans. *PLoS One* 2008;**3**:e1912. https://doi.org/10.1371/journal.pone.0001912

Michalettou T-D, Michalopoulos I, Costes SV *et al.* A meta-analysis of the effects of high-let ionizing radiations in human gene expression. *Life (Basel)* 2021;**11**:115. https://doi.org/10.3390/life11020115

Risso D, Ngai J, Speed TP *et al.* Normalization of rna-seq data using factor analysis of control genes or samples. *Nat Biotechnol* 2014;**32**:896–902. https://doi.org/10.1038/nbt.2931

Ritchie ME, Phipson B, Wu D *et al.* Limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Res* 2015;**43**:e47. https://doi.org/10.1093/nar/gkv007

Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;**26**:139–40. https://doi.org/10.1093/bioinformatics/btp616

Ron E. Ionizing radiation and cancer risk: evidence from epidemiology. *Pediatr Radiol* 2002;**32**:232. https://doi.org/10.1007/s00247-002-0672-0

Rothkamm K, Beinke C, Romm H *et al.* Comparison of established and emerging biodosimetry assays. *Radiat Res* 2013;**180**:111–9. https://doi.org/10.1667/rr3231.1

Sagkrioti E, Biz GM, Takan I *et al.* Radiation type- and dose-specific transcriptional responses across healthy and diseased mammalian tissues. *Antioxidants (Basel)* 2022;**11**:2286. https://doi.org/10.3390/antiox11112286

Salah A, Wollschläger D, Callari M *et al.* Genome-wide transcriptomic response of whole blood to radiation. *Sci Rep* 2025;**15**:19840. https://doi.org/10.1038/s41598-025-04898-1

Talapko J, Talapko D, Katalinić D *et al.* Health effects of ionizing radiation on the human body. *Medicina (Kaunas)* 2024;**60**:653. https://doi.org/10.3390/medicina60040653

Wilkinson MD, Dumontier M, Aalbersberg IJ *et al.* The fair guiding principles for scientific data management and stewardship. *Sci Data* 2016;**3**:160018. https://doi.org/10.1038/sdata.2016.18

Zhang Y, Parmigiani G, Johnson WE. Combat-seq: batch effect adjustment for rna-seq count data. *NAR Genom Bioinform* 2020;**2**:lqaa078. https://doi.org/10.1093/nargab/lqaa078

Bioinformatics Advances, 2026, 6, 1–5
https://doi.org/10.1093/bioadv/vbag092
Application Note