Prospects and challenges for FAIR toxicogenomics data

Saarimaki, Laura A.

2022-01

Saarimaki , L A , Melagraki , G , Afantitis , A , Lynch , I & Greco , D 2022 , ' Prospects and challenges for FAIR toxicogenomics data ' , Nature Nanotechnology , vol. 17 , no. 1 , pp. 17-18 . https://doi.org/10.1038/s41565-021-01049-1

http://hdl.handle.net/10138/355774
https://doi.org/10.1038/s41565-021-01049-1

acceptedVersion

nature nanotechnology

MATTERS ARISING

https://doi.org/10.1038/s41565-021-01049-1

Check for updates

Prospects and challenges for FAIR toxicogenomics data

Laura A. Saarimäki[1], Georgia Melagraki[2], Antreas Afantitis[2], Iseult Lynch[3] and Dario Greco[1,4] ✉

ARISING FROM Nina Jeliazkova et al. *Nature Nanotechnology* https://doi.org/10.1038/s41565-021-00911-6 (2021)

The article by Jeliazkova at al.[1] recently published in this journal addresses the pivotal topic of data sharing and reuse in nanosafety. Current research in the field is highly multidisciplinary, as described also in the recent call for reporting standards for bio–nano experimental studies[2]. Hence, the application of the general FAIR (findable, accessible, interoperable and reusable) principles[3], although valid, might fall short when considering field-specific needs and requirements. This is especially true for toxicogenomics, in which additional challenges are posed by the articulated data analytics as well as the need to integrate multiple datasets to increase the statistical power and domain of applicability of the resulting predictive models. These limitations substantially affect the possibility of including toxicogenomics-based evidence in safe-by-design protocols as well as in regulatory hazard and risk decisions.

In our recent effort to curate publicly available transcriptomics data from exposures with engineered nanomaterials[4], we initially identified 124 datasets. However, although nearly all these datasets were published in peer-reviewed articles, the data quality assessment resulted in the exclusion of 35 datasets due to problems in their overall usability, rather than reusability. These problems were primarily related to the experimental design, which suggests that several toxicogenomics datasets published in peer-reviewed articles present substantial design flaws that jeopardize the validity of any results extrapolated from them and stresses the need to critically evaluate even data that have been FAIRified. In other words, reinforcing rigorous reporting of data does not automatically ensure quality, which should be addressed in the early phases of the experimental design. In fact, our curation also raised another concern: even datasets deposited in established databases could still be made (more) FAIR[5] as, despite the availability of mature standards for minimum reporting of omics experiments (for example, MIAME[6] and MINSEQE (http://fged.org/projects/minseqe/)) to aid data FAIRness, several aspects remain undocumented in toxicogenomics studies. According to community-accepted minimum reporting standards and the FAIR principles, the primary experimental variables are to be described (for example, exposure doses and times). However, when it comes to the preprocessing and analysis of toxicogenomics data, these minimum standards often result in poor (re) usability due to the lack of batch-effect description (that is, potential systematic effects caused by reagents, microarrays and so on)[7–9] and incomplete characterization of the experimental design and execution[7]. This, in turn, prevents optimal data preprocessing and analysis, but could be easily overcome through additional criteria and quality checks built into the study design and reported as part of the required metadata.

Moreover, the reliance on minimum standards over complete documentation is not just a concern for the reuse of raw omics data. Similar challenges exist regarding the analysis and modelling performed on these data, which include the identification of predictive biomarkers, the development of adverse outcome pathways or the performance of the meta-analysis. Although the complexity of toxicogenomics data requires the use of articulated multistep analytical pipelines, their high dimensionality dictates the tailoring of algorithms and parameters to fit the specific characteristics of each experimental design and dataset. This has a profound impact, as equally technically valid alternative analytical strategies can lead to apparently divergent sets of results. Omics data analysis traditionally results in long lists of molecules that distinguish the experimental conditions assayed. These are intrinsically difficult to interpret unless functional analysis is performed to pinpoint over-represented biological functions. As the association of individual molecules with biological functions is, per se, an interpretative exercise, it is intuitive that alternative analytical strategies, which may result in slightly different sets of candidate molecules, may have a considerable impact on the interpretation of the final outcome. Indeed, this is one of the main reasons why toxicogenomics data still struggles to be fully accepted for regulatory purposes. Thus, ensuring the FAIRness of the computational protocols, tools and algorithms used to analyze toxicogenomics data can provide a sensible way to alleviate this bottleneck. In this regard, we advocate the need to differentiate between technical and scientific FAIRness[10]. Although the former can be addressed by sharing code, scripts and software to replicate a specific analysis, the latter focuses on the generation and sharing of standard operating procedures in which each analytical step is carefully motivated and described (metadata). Both technical and scientific FAIRness are equally important, albeit with slightly different 'owners' responsible for their implementation, and as a community we should define specific scientific FAIR principles for each of the different subdomains of nanosafety.

Finally, data curation is needed to advance research in many fields of modern science, and recognition of this huge effort is essential. Acknowledgement of the data generation effort is easily achieved through the publication of original research articles. However, curation of already published data often remains a sterile exercise in which the curated data, with increased FAIRness scores, remain fully available only to a small community of scientists. We propose two solutions to be adopted by authors and publishers, respectively. The former should consider curation as a valuable contribution to the field, and as such should publish the curated dataset and the associated curation protocols in one of

[1]Finnish Hub for Development and Validation of Integrated Approaches (FHAIVE), Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland. [2]Nanoinformatics Department, NovaMechanics Ltd, Nicosia, Cyprus. [3]School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham, UK. [4]Institute of Biotechnology, University of Helsinki, Helsinki, Finland. ✉e-mail: dario.greco@tuni.fi

the myriad of data-focused journals. Publishers can contribute by requiring the bulk of the curated data that underpins meta-analyses and chemo- and nanoinformatics models to be accessible via well-established data repositories (such as Zenodo), via specific open curation databases (for example, the NanoPharos Database (https://db.nanopharos.eu/Queries/Datasets.zul)) and/or via other database platforms. Reuse of curated data will be facilitated by ensuring that the data are exported in formats that are suitable for modelling or further analysis.

With these considerations in mind, we believe that it is meaningful to address the overall usability of published data in addition to the aspects of FAIR, and that the usability can be improved through many of the actions already suggested by the nanosafety community[1,2,5,7–11]. The challenges discussed in this comment are not unique to nanosafety but pervade the toxicogenomics field as a whole. However, notable efforts, such as that by Jeliazkova et al.[1], place the nanosafety community at the forefront of advancing the entire area of chemical safety assessment. Indeed, the nanosafety community is driving the updating of regulatory testing on a wide scale. Supplementing the broad technical FAIR principles with subdomain-specific considerations, as represented here by the toxicogenomics field, will considerably increase the transparency of results and predictions based on the reuse of such data. Furthermore, it will pave the way towards regulatory acceptance of toxicogenomics-based evidence in the safety assessment of engineered nanomaterials and other chemicals alike.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41565-021-01049-1.

Received: 26 May 2021; Accepted: 11 November 2021;
Published online: 23 December 2021

References

1. Jeliazkova, N. et al. Towards FAIR nanosafety data. *Nat. Nanotechnol.* **16**, 644–654 (2021).
2. Faria, M. et al. Minimum information reporting in bio–nano experimental literature. *Nat. Nanotechnol.* **13**, 777–785 (2018).
3. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
4. Saarimäki, L. A. et al. Manually curated transcriptomics data collection for toxicogenomic assessment of engineered nanomaterials. *Sci. Data* **8**, 49 (2021).
5. Ammar, A. et al. A semi-automated workflow for FAIR maturity indicators in the life sciences. *Nanomaterials* **10**, 2068 (2020).
6. Brazma, A. et al. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nat. Genet.* **29**, 365–371 (2001).
7. Kinaret, P. A. S. et al. Transcriptomics in toxicogenomics, part I: experimental design, technologies, publicly available data, and regulatory aspects. *Nanomaterials* **10**, 750 (2020).
8. Federico, A. et al. Transcriptomics in toxicogenomics, part II: preprocessing and differential expression analysis for high quality data. *Nanomaterials* **10**, 903 (2020).
9. Mülhopt, S. et al. Characterization of nanoparticle batch-to-batch variability. *Nanomaterials* **8**, 311 (2018).
10. Papadiamantis, A. G. et al. Metadata stewardship in nanosafety research: community-driven organisation of metadata schemas to support FAIR nanoscience data. *Nanomaterials* **10**, 2033 (2020).
11. Serra, A. et al. Transcriptomics in toxicogenomics, part III: data modelling for risk assessment. *Nanomaterials* **10**, 708 (2020).

Acknowledgements
We acknowledge funding from the European Union's Horizon 2020 research and innovation programme via the NanoSolveIT Project (grant agreement no. 814572) and the Academy of Finland (grant agreement no. 322761).

Author contributions
L.A.S. carried out the formal analysis, contributed to the data curation and methodology, and co-wrote the original draft. G.M. contributed to the methodology, review and editing. A.A. contributed to the methodology, review and editing, and funding acquisition. I.L. contributed to the data curation and funding acquisition, and co-wrote the original draft. D.G. conceptualized and supervised the work, contributed to the funding acquisition and co-wrote the original draft.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to Dario Greco.