D1502–D1506 *Nucleic Acids Research, 2021, Vol. 49, Database issue* *Published online 19 November 2020*
doi: 10.1093/nar/gkaa1062

From ArrayExpress to BioStudies

Ugis Sarkans, Anja Füllgrabe, Ahmed Ali, Awais Athar, Ehsan Behrangi, Nestor Diaz, Silvie Fexova, Nancy George, Haider Iqbal, Sandeep Kurri, Jhoan Munoz, Juan Rada, Irene Papatheodorou and Alvis Brazma*

European Molecular Biology Laboratory, European Bioinformatics Institute, EMBL-EBI, Wellcome Trust Genome Campus, Hinxton CB10 1SD, UK

Received September 21, 2020; Revised October 16, 2020; Editorial Decision October 19, 2020; Accepted October 27, 2020

ABSTRACT

ArrayExpress (https://www.ebi.ac.uk/arrayexpress) is an archive of functional genomics data at EMBL-EBI, established in 2002, initially as an archive for publication-related microarray data and was later extended to accept sequencing-based data. Over the last decade an increasing share of biological experiments involve multiple technologies assaying different biological modalities, such as epigenetics, and RNA and protein expression, and thus the BioStudies database (https://www.ebi.ac.uk/biostudies) was established to deal with such multimodal data. Its central concept is a *study*, which typically is associated with a publication. BioStudies stores metadata describing the study, provides links to the relevant databases, such as European Nucleotide Archive (ENA), as well as hosts the types of data for which specialized databases do not exist. With BioStudies now fully functional, we are able to further harmonize the archival data infrastructure at EMBL-EBI, and ArrayExpress is being migrated to BioStudies. In future, all functional genomics data will be archived at BioStudies. The process will be seamless for the users, who will continue to submit data using the online tool Annotare and will be able to query and download data largely in the same manner as before. Nevertheless, some technical aspects, particularly programmatic access, will change. This update guides the users through these changes.

INTRODUCTION

ArrayExpress is an archive of functional genomics data, such as gene expression or DNA methylation profiling data (1). ArrayExpress is the main source of data for Expression Atlas (2) – an added-value gene expression database at EMBL-EBI, which allows for gene-, tissue- or disease-based queries. ArrayExpress was established as an archive for microarray data in 2002 (3) as the first MIAME-compliant public database (4). With technology evolving, in 2008 ArrayExpress was extended to accept sequencing-based functional genomics data, in particular, data from RNA sequencing assays (5). For these experiments ArrayExpress stores the processed data and experimental metadata, brokering the sequences to the European Nucleotide Archive (ENA) (6). Starting from 2017, the volume of submissions from sequencing-based experiments has exceeded those from microarrays. For selected transcriptomics experiments, the data are consistently re-processed and re-annotated by our curation team and are made available in Expression Atlas. A major shift over the last two years has been a rapid increase in data from experiments providing cell-level resolution, namely single-cell RNA-seq experiments (7). ArrayExpress is one of the Core Data Resources of the European bioinformatics infrastructure ELIXIR since 2017 (8). Data is submitted to ArrayExpress via the submission tool Annotare (9), which outputs standardized MAGE-TAB format files (10), which are then loaded into ArrayExpress.

Molecular biology experiments are becoming increasingly multimodal and often employ a range of technologies, for example combining RNA-seq, protein expression assays and genotyping. To deal with data from such multimodal experiments, the BioStudies database (www.ebi.ac.uk/biostudies) was established at EMBL-EBI in 2016 (11,12). Its central concept is a *study*, which typically is associated with a publication. BioStudies stores metadata describing the study, provides links to the relevant databases, such as ENA for sequencing or PRIDE for proteomics experiments (13), and also stores the actual data from technologies for which specialized databases do not exist (for instance, microscopy). Thus, BioStudies provides a means to package all the data associated with a (peer-reviewed) publication and provides a more flexible way to organize the data than ArrayExpress. The overarching goal of BioStudies is to support transparency and reproducibility of life sciences research by aggregating all the outputs of a study in a single place. BioStudies also includes the concept of a *collection* of studies, which allows for grouping of datasets that

*To whom correspondence should be addressed. Tel: +44 1223 494 658; Fax: +44 1223 494 468; Email: brazma@ebi.ac.uk

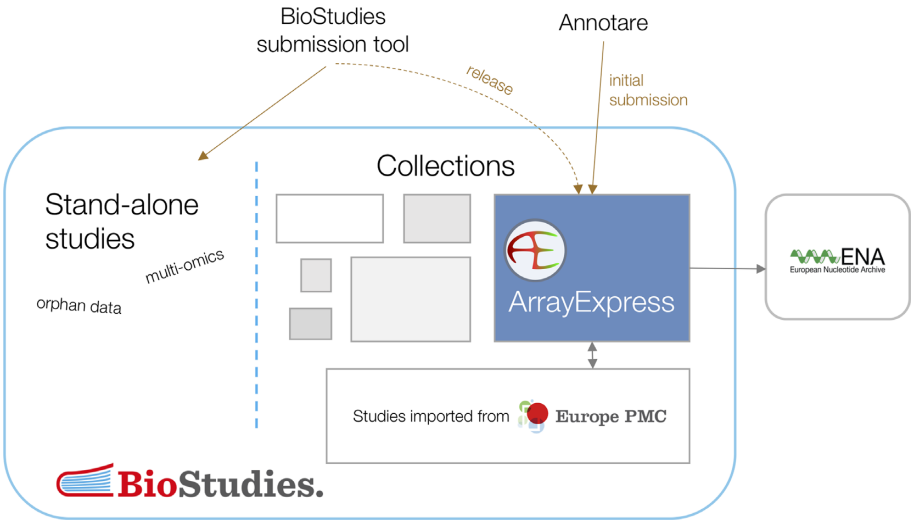

Figure 1. Functional genomics data will be migrated to the ArrayExpress collection in BioStudies and new submissions will be loaded via Annotare into BioStudies (with sequencing data being brokered to ENA).

share particular common features or have been acquired from a single source, for example a collection of toxicogenomics datasets gathered in the diXa project (14). An example of linking data from different modalities used in the same study, is study S-BSST390 that contains serial electron microscopy images, links to EMDB entries, a link to the PRIDE database, and a link to another entry within BioStudies.

It should be noted that currently the majority of BioStudies records are created after the related paper has been published and a pipeline from Europe PMC (15) is the main source of data in BioStudies. In fact, a BioStudies record is available for all open articles in Europe PMC that have auto-detected links to life sciences databases, supplementary materials, or both. Other data sources are the SourceData project (16) and image datasets associated with papers in the Journal of Cell Biology. However, we are increasingly focusing on pre-publication data submissions, which allow for citing the respective BioStudies record. Since 2017, BioStudies is an ELIXIR Deposition Database (https://elixir-europe.org/platforms/data/elixir-deposition-databases) and the number of direct submissions is increasing rapidly.

Being more general and designed with multimodal studies in mind, the BioStudies database will supersede ArrayExpress in 2021. All existing ArrayExpress data will be made available unchanged in BioStudies as a part of the ArrayExpress data collection and all accession numbers will be retained (Figure 1). The BioStudies data search, exploration and API provide all the current ArrayExpress functionality and will be further tuned in response to community feedback. Thus, in the future, all functional genomics data and metadata will be archived in BioStudies. The process will be mostly seamless for the users, who will continue to submit data using Annotare as before and will be able to query and download the data largely in the same manner as currently in ArrayExpress. Nevertheless, some technical aspects, particularly programmatic access, will change, as further detailed below alongside other main developments.

DATA SUBMISSIONS AND GROWTH

ArrayExpress has continued to grow in data content, and new templates have been added to the Annotare submission tool to improve the user experience, mainly with respect to single-cell RNA-sequencing data.

Annotare developments

The single-cell template (Figure 2) provides the submitter with a selection of single-cell specific sample attributes and a new tab to include the details about the single-cell library construction. Here, the user may pick from a list of the most commonly used single-cell protocols, and library attributes, such as 'end bias' or 'UMI barcode size', will be prefilled with the default values for the selected protocol.

Performance was improved when importing data files after FTP/Aspera upload, and a few changes have improved the user interface, e.g. upload of erroneous file formats is prevented, and useful validation checks and warnings have been added. The submission guide and help content that were previously split between ArrayExpress and Annotare home pages have been integrated into Annotare's website. The submission help has also been updated to the latest EMBL-EBI style framework for a uniform look and feel.

Submission statistics and trends

Since 2017, there have been on average 1000 newly submitted experiments per year. In 2020, single-cell RNA-seq experiments make up more than 10% of the submitted functional genomics experiments while microarray submissions continue to decline (Figure 3). ArrayExpress has recently received its first submission of data produced via the CITE-Seq protocol (17), which generates RNA-seq and protein expression data of cell surface proteins and antibodies in parallel at the single-cell level. Similarly, we have started handling a limited number of spatially resolved transcriptomics protocols, generated in particular via the Visium

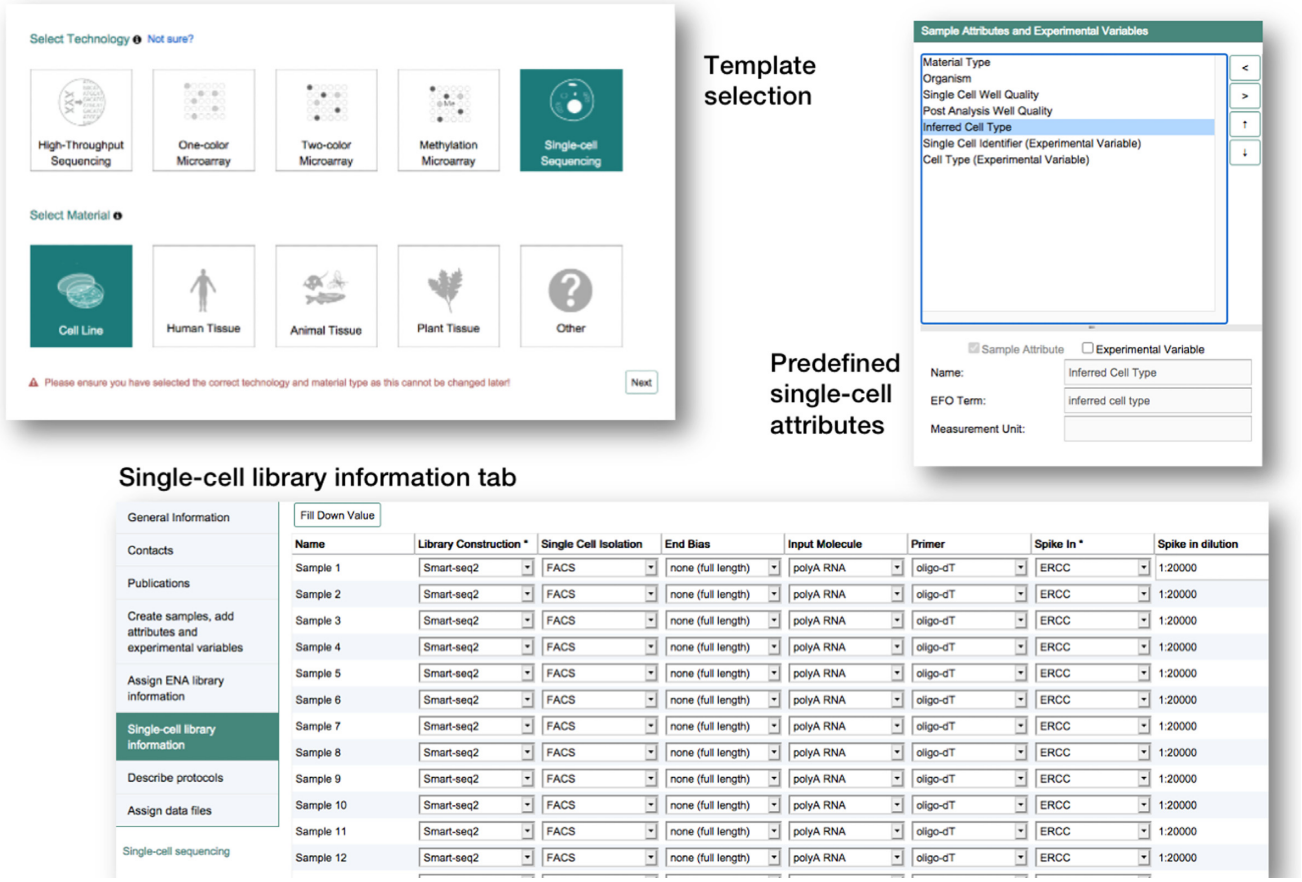

Figure 2. A refined template selection and fields to capture single-cell specific attributes have been added to Annotare.

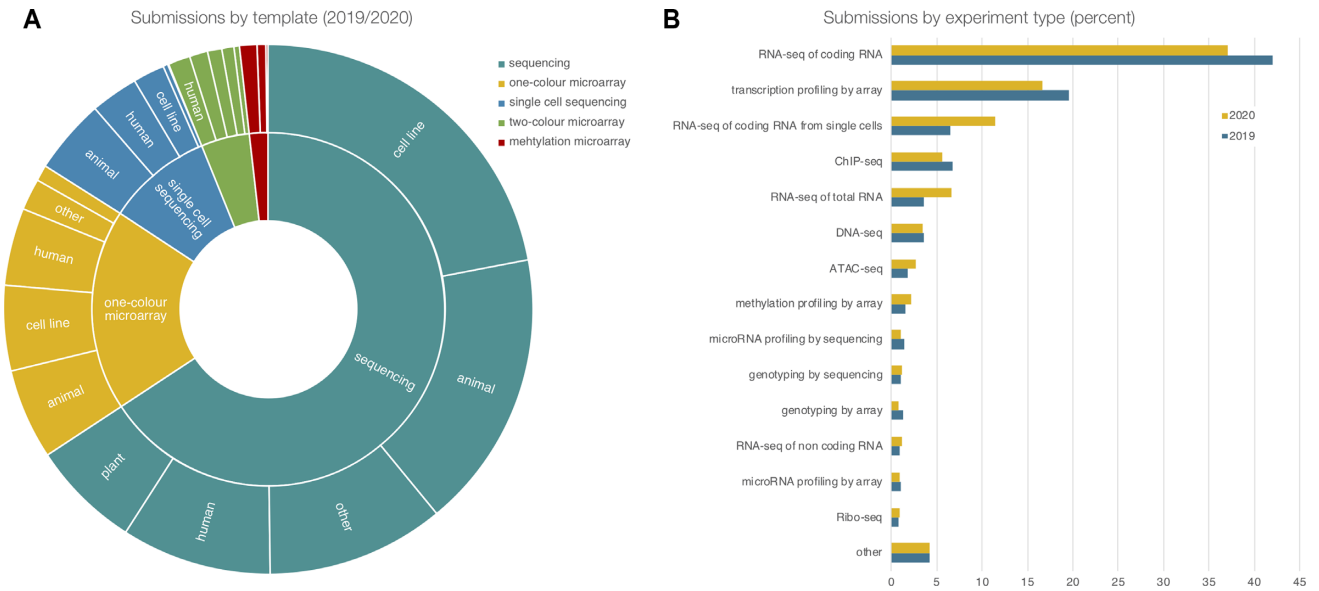

Figure 3. Experiment submissions via Annotare from January 2019 to September 2020; (**A**) broken down by template type, which is composed of technology (inner ring) and biological starting material (outer ring); and (**B**) by experiment type, showing the increase of single-cell RNA-seq (note that 'RNA-seq of total RNA' was added new in May 2019).

platform, indicating a trend toward multi-omic and spatial data submissions.

ACCOMMODATING ARRAYEXPRESS DATA IN BIOSTUDIES

The BioStudies database has been described previously (12), however several new features have been developed, largely to provide the BioStudies ArrayExpress data collection with functionality equivalent to that currently available in ArrayExpress. In particular, our file access interface component can now work with large datasets (millions of files in the same study) and for large data volumes we support FTP and Aspera protocols, both for data depositions and access. A dataset can stay private, i.e. visible only to authorized users, while the associated publication goes through the peer-review process. We also provide a means to filter data via a generic faceting mechanism, which can be easily tuned for the needs of a specific data collection within BioStudies, for easy data exploration. For the ArrayExpress data collection the browser includes facets such as 'Organism', 'Study type' and 'Technology type'.

Data and data representation

While in ArrayExpress the main unit of information is an *experiment*, in BioStudies it is a *study*. These are comparable concepts, and each of the existing ArrayExpress experiments is being translated into a separate study in BioStudies. As already noted, ArrayExpress data entries will retain the accession numbers in BioStudies; moreover, the existing hyperlinks to ArrayExpress will be redirected to BioStudies. All new datasets acquired via Annotare will be loaded into BioStudies, initially in parallel with loading into ArrayExpress, and then exclusively.

The ArrayExpress approach to representing a dataset is based on visualizing the sample-to-data relationship table, which is a part of the MAGE-TAB format (10). BioStudies retains the original MAGE-TAB files and provides a widget for browsing the MAGE-TAB samples table. In addition, utilizing the generic nature of the BioStudies data model we offer other methods to explore a dataset. For example, for each dataset we show a summary table of all unique combinations of experimental factor values, together with the number of samples for these combinations and quick access to the relevant data files. Adding or refining various data representation aspects in response to community feedback is easy to achieve in the BioStudies framework.

API access

The existing ArrayExpress Application Programmatic Interface (API) functionality will be available from BioStudies, though some changes will happen. API responses in XML format are being deprecated in favor of JSON. However, XML representation is available for individual datasets in a more general schema. We encourage users to shift to the new JSON schema which encapsulates all the information available in the current ArrayExpress JSON. The search API will return a paged response containing only limited metadata about the list of studies being retrieved. Complete metadata on individual studies will not be bundled together, and users will need to iterate through the set and request each study individually. The endpoint changes from https://www.ebi.ac.uk/arrayexpress/xml/v3 to https://www.ebi.ac.uk/biostudies/api/v1. A list of all searchable fields and their BioStudies equivalent is provided in the migration guide available as Supplementary Note and via https://www.ebi.ac.uk/biostudies/ArrayExpress/help. Additionally, the default BioStudies search criteria are available as well.

Post-submission modifications

ArrayExpress allows specific types of post-submission modifications of the database record to be performed by the submitter, such as changing the public release date or adding a publication reference. Submitting these modifications will be made possible via BioStudies.

Stages and timeline

The ArrayExpress migration will take place in two phases. During the first phase, scheduled to start in October 2020, all new datasets will be available both from the current ArrayExpress and BioStudies resources. The existing ArrayExpress datasets will become available via BioStudies and the user feedback will be collected. In the second phase, by summer 2021, the current ArrayExpress infrastructure will be deprecated and all data will become a part of the ArrayExpress data collection in BioStudies.

FUTURE PLANS

The ArrayExpress to BioStudies process is still on-going. For instance, currently Annotare and BioStudies have separate user accounts for user authorization. This will be addressed by supporting the ELIXIR AAI system (18), which will allow for a single login to all EMBL-EBI data resources. EMBL-EBI maintains a range of databases for specific types of life science data and knowledge (19), most with custom-built data deposition. The submissions of multimodal experiments to the EMBL-EBI resources is a more general question, that goes beyond the remit of this ArrayExpress update and here we discuss the relevant future developments only briefly.

The BioStudies future developments directly relevant to ArrayExpress data submitters are related to Annotare. The development of Annotare will focus on improvements for more efficient submission throughput and handling of larger submissions. As both the number of submissions and their respective size increase (in terms of sample numbers and data volume), faster file import, validation and submission processing will be developed. Changes on the interface will be implemented for easier annotation of large numbers of samples. We will also continue to adjust the templates to the possible shifts in technologies. As techniques such as CITE-Seq and spatially resolved transcriptomics become more commonly used, we will expect to make updates in the MAGE-TAB representation of metadata for these types of studies.

For well-established types of data, such as gene expression or protein mass spectrometry, a custom-built data de-

position tool offers the best support for their users and ensures the reusability of the collected data. Typical examples of such tools are Annotare and the PRIDE proteomics data submission tool (13). However, a more generic data acquisition system is needed for data generated by new technologies during the early stages of their development, e.g. currently for light microscopy (20). To serve both needs, BioStudies offers two submission systems: Annotare for gene expression and functional genomics data collected to the standard that enables population of Expression Atlas, and a more generic data submission tool for types of data for which specialized resources do not yet exist. Unfortunately, this means that currently to submit data from multimodal experiments to the EMBL-EBI resources, several separate submissions may be needed, which then can be linked via BioStudies (the linking happens automatically for data related to publications available in Europe PMC). The roadmap toward simplifying this process will be described in the next BioStudies update. The migration of ArrayExpress to BioStudies is part of our work to simplify EMBL-EBI's data resource ecosystem.

This is the last ArrayExpress update in the Nucleic Acid Research Journal Database Issue since future updates will be a part of the BioStudies database and Expression Atlas updates. However, we would like to emphasize that all the data submitted to ArrayExpress since 2002 will be available from BioStudies database without any changes and the accession numbers will be preserved. The ArrayExpress migration to BioStudies reflects the changing field of life sciences, where multimodal high-throughput experiments have become a norm.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Single Cell Gene Expression Atlas grant from the Wellcome Trust [108437/Z/15/Z]. Funding for open access charge: EMBL-EBI.
Conflict of interest statement. None declared.

REFERENCES

1. Athar,A., Fullgrabe,A., George,N., Iqbal,H., Huerta,L., Ali,A., Snow,C., Fonseca,N.A., Petryszak,R., Papatheodorou,I. *et al.* (2019) ArrayExpress update - from bulk to single-cell expression data. *Nucleic Acids Res.*, **47**, D711–D715.
2. Papatheodorou,I., Moreno,P., Manning,J., Fuentes,A.M., George,N., Fexova,S., Fonseca,N.A., Füllgrabe,A., Green,M., Huang,N. *et al.* (2020) Expression Atlas update: from tissues to single cells. *Nucleic Acids Res.*, **48**, D77–D83.
3. Brazma,A., Parkinson,H., Sarkans,U., Shojatalab,M., Vilo,J., Abeygunawardena,N., Holloway,E., Kapushesky,M., Kemmeren,P., Lara,G.G. *et al.* (2003) ArrayExpress–a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.*, **31**, 68–71.
4. Brazma,A., Hingamp,P., Quackenbush,J., Sherlock,G., Spellman,P., Stoeckert,C., Aach,J., Ansorge,W., Ball,C.A. and Causton,H.C. (2001) Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nat. Genet.*, **29**, 365–371.
5. Parkinson,H., Sarkans,U., Kolesnikov,N., Abeygunawardena,N., Burdett,T., Dylag,M., Emam,I., Farne,A., Hastings,E., Holloway,E. *et al.* (2011) ArrayExpress update–an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res.*, **39**, D1002–D1004.
6. Amid,C., Alako,B.T.F., Balavenkataraman Kadhirvelu,V., Burdett,T., Burgin,J., Fan,J., Harrison,P.W., Holt,S., Hussein,A., Ivanov,E. *et al.* (2020) The European Nucleotide Archive in 2019. *Nucleic Acids Res.*, **48**, D70–D76.
7. Wang,Z., Gerstein,M. and Snyder,M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.
8. Drysdale,R., Cook,C.E., Petryszak,R., Baillie-Gerritsen,V., Barlow,M., Gasteiger,E., Gruhl,F., Haas,J., Lanfear,J., Lopez,R. *et al.* (2020) The ELIXIR core data resources: fundamental infrastructure for the life sciences. *Bioinformatics*, **36**, 2636–2642.
9. Kolesnikov,N., Hastings,E., Keays,M., Melnichuk,O., Tang,Y.A., Williams,E., Dylag,M., Kurbatova,N., Brandizi,M., Burdett,T. *et al.* (2015) ArrayExpress update–simplifying data submissions. *Nucleic Acids Res.*, **43**, D1113–D1116.
10. Rayner,T.F., Rocca-Serra,P., Spellman,P.T., Causton,H.C., Farne,A., Holloway,E., Irizarry,R.A., Liu,J., Maier,D.S., Miller,M. *et al.* (2006) A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. *BMC Bioinformatics*, **7**, 489.
11. McEntyre,J., Sarkans,U. and Brazma,A. (2015) The BioStudies database. *Mol. Syst. Biol.*, **11**, 847.
12. Sarkans,U., Gostev,M., Athar,A., Behrangi,E., Melnichuk,O., Ali,A., Minguet,J., Rada,J.C., Snow,C., Tikhonov,A. *et al.* (2018) The BioStudies database-one stop shop for all data supporting a life sciences study. *Nucleic Acids Res.*, **46**, D1266–D1270.
13. Perez-Riverol,Y., Csordas,A., Bai,J., Bernal-Llinares,M., Hewapathirana,S., Kundu,D.J., Inuganti,A., Griss,J., Mayer,G., Eisenacher,M. *et al.* (2019) The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.*, **47**, D442–D450.
14. Hendrickx,D.M., Aerts,H.J., Caiment,F., Clark,D., Ebbels,T.M., Evelo,C.T., Gmuender,H., Hebels,D.G., Herwig,R., Hescheler,J. *et al.* (2015) diXa: a data infrastructure for chemical safety assessment. *Bioinformatics*, **31**, 1505–1507.
15. Levchenko,M., Gou,Y., Graef,F., Hamelers,A., Huang,Z., Ide-Smith,M., Iyer,A., Kilian,O., Katuri,J., Kim,J.H. *et al.* (2018) Europe PMC in 2017. *Nucleic Acids Res.*, **46**, D1254–D1260.
16. Liechti,R., George,N., Götz,L., El-Gebali,S., Chasapi,A., Crespo,I., Xenarios,I. and Lemberger,T. (2017) SourceData: a semantic platform for curating and searching figures. *Nat. Methods*, **14**, 1021–1022.
17. Stoeckius,M., Hafemeister,C., Stephenson,W., Houck-Loomis,B., Chattopadhyay,P.K., Swerdlow,H., Satija,R. and Smibert,P. (2017) Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods*, **14**, 865–868.
18. Linden,M., Prochazka,M., Lappalainen,I., Bucik,D., Vyskocil,P., Kuba,M., Silén,S., Belmann,P., Sczyrba,A., Newhouse,S. *et al.* (2018) Common ELIXIR service for researcher authentication and authorisation. *F1000Res*, **7**, ELIXIR-1199.
19. Cook,C.E., Stroe,O., Cochrane,G., Birney,E. and Apweiler,R. (2020) The European Bioinformatics Institute in 2020: building a global infrastructure of interconnected data resources for the life sciences. *Nucleic Acids Res.*, **48**, D17–D23.
20. Ellenberg,J., Swedlow,J.R., Barlow,M., Cook,C.E., Sarkans,U., Patwardhan,A., Brazma,A. and Birney,E. (2018) A call for public archives for biological image data. *Nat. Methods*, **15**, 849–854.