TYPE Review
PUBLISHED 20 January 2025
DOI 10.3389/fimmu.2024.1439434

Check for updates

OPEN ACCESS

EDITED BY
Enrique Hernandez-Lemus,
National Institute of Genomic Medicine (INMEGEN), Mexico

REVIEWED BY
C Gopi Mohan FRSC,
Amrita Vishwa Vidyapeetham University, India
Guillermo De Anda Jáuregui,
National Institute of Genomic Medicine (INMEGEN), Mexico

*CORRESPONDENCE
Douaa Mugahid
✉ dmugahid@hsph.harvard.edu

RECEIVED 27 May 2024
ACCEPTED 17 December 2024
PUBLISHED 20 January 2025

CITATION
Mugahid D, Lyon J, Demurjian C, Eolin N, Whittaker C, Godek M, Lauffenburger D, Fortune S and Levine S (2025) A practical guide to FAIR data management in the age of multi-OMICS and AI. *Front. Immunol.* 15:1439434. doi: 10.3389/fimmu.2024.1439434

A practical guide to FAIR data management in the age of multi-OMICS and AI

Douaa Mugahid[1]*, Jared Lyon[2], Charlie Demurjian[2], Nathan Eolin[2], Charlie Whittaker[2], Mark Godek[3], Douglas Lauffenburger[4], Sarah Fortune[1] and Stuart Levine[2]

[1]Department of Immunology and Infectious Diseases, T.H. Chan School of Public Health, Harvard University, Boston, MA, United States, [2]BioMicro Center, Massachusetts Institute of Technology, Cambridge, MA, United States, [3]Ragon Institute of Massachusetts General Hospital (MGH), Massachusetts Institute of Technology (MIT), and Harvard, Cambridge, MA, United States, [4]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, United States

Multi-cellular biological systems, including the immune system, are highly complex, dynamic, and adaptable. Systems biologists aim to understand such complexity at a quantitative level. However, these ambitious efforts are often limited by access to a variety of high-density intra-, extra- and multi-cellular measurements resolved in time and space and across a variety of perturbations. The advent of automation, OMICs and single-cell technologies now allows high dimensional multi-modal data acquisition from the same biological samples multiplexed at scale (multi-OMICs). As a result, systems biologists -theoretically- have access to more data than ever. However, the mathematical frameworks and computational tools needed to analyze and interpret such data are often still nascent, limiting the biological insights that can be obtained without years of computational method development and validation. More pressingly, much of the data sits in silos in formats that are incomprehensible to other scientists or machines limiting its value to the vaster scientific community, especially the computational biologists tasked with analyzing these vast amounts of data in more nuanced ways. With the rapid development and increasing interest in using artificial intelligence (AI) for the life sciences, improving how biologic data is organized and shared is more pressing than ever for scientific progress. Here, we outline a practical approach to multi-modal data management and FAIR sharing, which are in line with the latest US and EU funders' data sharing policies. This framework can help extend the longevity and utility of data by allowing facile use and reuse, accelerating scientific discovery in the biomedical sciences.

KEYWORDS
FAIR data, systems biology, immunology, OMICs, multi-modal data, artificial intelligence, modeling, Science administration

Introduction

Data powers our understanding of the world around us. As the world becomes fully digitized and technology continues to develop, researchers' ability to gather different types of measurements at scale is only increasing, making the adoption of Data Science principles across all disciplines increasingly necessary. This is particularly true in biomedical research, where the race to understand the basis of life and human disease has encouraged researchers to push the boundaries of method development for decades. Most notably, the advent of high-throughput technologies such as high-content imaging, multi-parameter flow cytometry/CyToF, microarrays, next generation sequencing, and mass spectrometry has transformed biologic research into a rich multi-modal data science.

The increase in scale across the research enterprise requires careful experimental design as well as the development of novel statistical and mathematical frameworks that can help researchers synthesize all this data into meaningful, and occasionally non-intuitive biological insights. We collectively refer to these frameworks as artificial intelligence (AI). Practically, this necessitates making the data accessible and interpretable to other researchers as well as machines in ways that allow it to be used for applications beyond its original intent. In the case of human studies, this also requires deliberate efforts to ensure the data is representative of human diversity and is well-guarded to protect individuals' privacy and rights. Only then can society fully benefit from the richness and complexity of the data needed to enable AI-driven advances in the biomedical field for the benefit of all.

A term commonly used to refer to good data stewardship in the life sciences is FAIR data sharing (1). The term is an acronym for data that is findable, accessible, interoperable, and reusable, all features that extend the usability of data beyond the purposes it was generated for, thereby increasing its long-term impact. Here, we outline practical steps towards FAIR data-sharing practices that can improve the longevity and utility of biomedical data. The principles are applicable to any field in the Life Sciences.

Why share data

Before ChatGPT made conversations about AI, multi-modal data, data and computational bias so mainstream (2, 3) the biomedical field had its taste of AI's enabling potential, yet relatively little light was shed on the central role community-wide data curation and sharing played in enabling these biomedical breakthroughs. We discuss two prominent examples below.

AlphaFold

In the case of AlphaFold (4), which won its developers the Lasker Award in 2023 (5) and Nobel Prize in 2024 (6), years of researchers publicly depositing experimentally determined protein structures and genomic sequences was fundamental to training the underlying model (4). These data resided in public databases such as UniRef90 (7), BFD (8), Uniclust30 (9), MGnify clusters (10),and the Protein Data Bank (PDB) (11) which house some of the world's largest collections of biologic sequences and structures, respectively. Since models are only as good as the data they were trained on, it is no surprise that proteins poorly predicted by AlphaFold are often classes that are underrepresented in nature either because they lack homologues such as orphan proteins (12), or because they are highly variable such as antibodies (13). Since its release, over a million AlphaFold predicted protein structures have been shared in the public domain (14), including the proteins of bacteria rapidly developing antibiotic resistance, thereby posing an urgent threat to global health. The entirety of the AlphaFold model is also available publicly for others to explore and expand (15), which has enabled researchers to adopt it for their use-case of choice expanding its impact even further (16). Without publicly deposited data the development of AlphaFold would not have been possible, neither would much of the science enabled by it.

NextStrain

As part of the global COVID19 response researchers worldwide rushed to sequence and share the SARS-CoV2 genomes they isolated. SARS-CoV2 genomic sequences were centralized in a repository called NextStrain (17), which also provided researchers the world over with tools that allowed tracking mutations in the viral genome, some of which threatened to be associated with changes in transmissibility, virulence, and/or clinical presentation thereby informing public health responses (18–21). Other researchers got straight to developing vaccines against the devastating virus (22). NextStrain's developers' focus on data and code sharing was central to its broad utility during the COVID19 pandemic, which was accompanied by an exponential increase in the number of citations from 19 in 2018 to over 2500 citations by May 2024. Its user-friendliness and transparency likely played a role in it featuring in several policy reports (18, 19, 21).

These two examples demonstrate how data and code sharing is fundamental to driving impactful scientific advances in the digital age. Both efforts required the development of globally accessible platforms that allowed the sharing of important, standardized biomedical data at scale. This enabled others to develop computational methods that could crunch through the massive volumes of data thereby providing more researchers with usable information on which to build. As more types of data are standardized and shared, we can only begin to imagine the scope and impact of future breakthroughs.

The shifts in funding agency requirements

While the utility of sharing SARS-CoV2 genomic data was likely obvious to many, it is difficult to imagine that the developers of PDB or NCBI predicted the development of AlphaFold. It's even more difficult to believe that every researcher depositing their protein structures since 1971 or sequences since 1982 understood they would be individual contributors to such a development many years down the line. Even if they did, it is no secret that the current scientific eco-system lacks a short-term mechanism for rewarding raw data sharing, which is a time-consuming and laborious process many researchers find unpleasant.

Incentive systems that reward data sharing are still under development but should be possible with the popularization of digital object identifiers (DOIs) (23), which allows users to uniquely cite papers as well as code and data in the digital sphere. In the meantime, science policy makers and funders seeking to maximize return on their and/or the public's investment in basic research have resorted to mandating data sharing. In the absence of clear mechanisms for accountability much of the enforcement currently falls to publishers. As a result, the emerging practice is for researchers to only share positive data or data that is included in a publication which means a lot of data remains unaccounted for. Another important implication is that AI algorithms are being trained disproportionately on positive data which will affect their performance and generalizability (24). That said, enforcement at publication has served science well and is arguably one of the main reasons PDB is now populated with close to 200,000 experimentally validated protein structures. Among the funders now encouraging data and code deposition are the NIH (25) and NSF (26) which are currently updating their policies following mandates from the Whitehouse Office of Science and Technology (27), as well as philanthropic organizations such as the Bill and Melinda Gates Foundation (28), and the Chan Zuckerberg Initiative (29) in the USA. In Europe, the Wellcome Trust (30) and Horizon Europe (31) mandate FAIR data sharing whenever possible.

Key elements for responsible data use and informed reuse

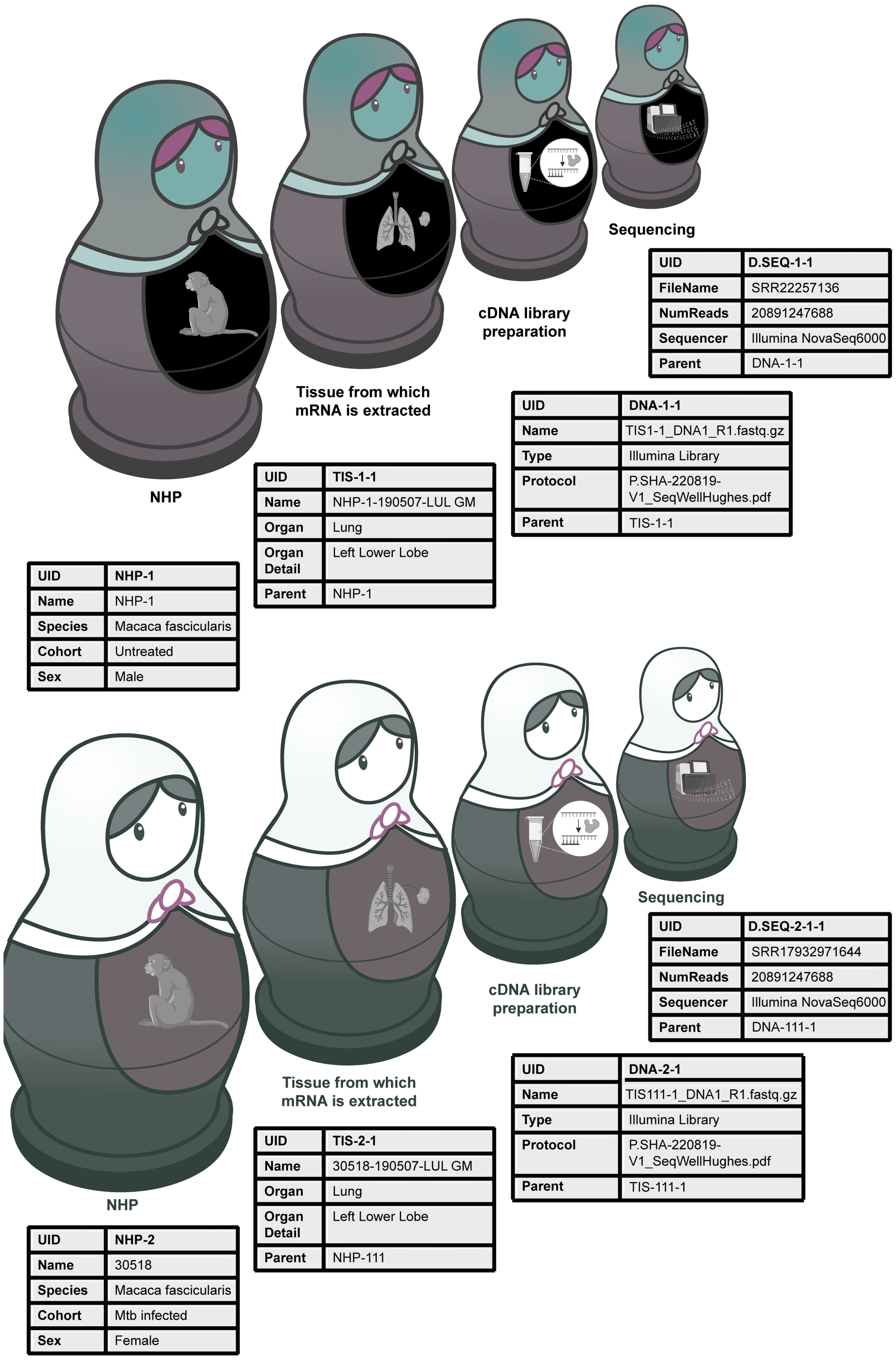
Rich metadata provides necessary context

Experimental data are most reusable when associated with rich metadata, often referred to as data about data, that help future users interpret and differentiate between data sets and individual data points. Imagine a set of hand-made Russian dolls. The smallest doll (the dataset) is nested within other dolls (the layers of metadata collected at each experimental step leading up to the data). Being hand-made, the innermost doll from a single set is made to fit the outer dolls perfectly, but might not fit within another set of Russian dolls, even if they were made by the same craftsperson, and even less so if made by another. If the craftsperson were to share the dimensions of each of the dolls, however, they could help others predict which inner dolls are combinable between sets (Figure 1). The more detailed and understandable the dimensions, the better the predicted fit, thus the need for rich and standardized metadata. In a research setting, capturing rich and understandable metadata is useful not only for researchers in the same lab, but also when re-used by others once the data is in the public domain, and is key to interoperability. Technical and biological confounders can often skew the interpretation of data and can only be accounted for if they are reported as part of the original study. This practice minimizes the chances that others will re-use the data under false assumptions leading to the generation of poorly informed hypotheses. The result is an avoidable loss of time, energy, and resources of many chasing the wrong ideas. More importantly, this contributes to the safe and cost-effective development of safer medicines for patients if such data is ever to be used in that context.

As an example, consider DNA sequencing data from a series of different tissue biopsies from a non-human primate. Each raw sequencing file is linked to a DNA library, prepared from a tissue, extracted from an animal. At the time the animal is taken into a study, it's important to assign it a unique identifier, and document its age, sex, species and geographical origins, any interventions it underwent and when, as well as the organ from which the sample was extracted, plus the time and method of extraction. Each biopsy should also receive a unique identifier and be linked to the parent animal. Similarly, the DNA library should also receive a unique identifier and be linked to the biopsy from which it was prepared, together with information about the DNA extraction kit/process (eg: the thermocycler used, the number of amplification cycles, the temperature at each step, and the sequencing primers). Finally, it's also important to note which samples were multiplexed on which chip (which should also have unique identifiers), the sequencer used, read length, and sequencing depth. If a plate-based fluorescence assay was done on cells from the same tissue, which sample was in which well, what was the analyte, antibody, and fluorophore, which plate reader was used, on which day, and what were the excitation and emission spectra. Much of this information can be captured in independent tables that serve as templates for researchers at every experimental step and can later be linked together as a set of relational databases, a simple but elegant and well-established solution in Data Science, that ensures full data provenance for single or multi-modal datasets from the same experiment. The metadata can then be shared in the public domain making the associated data truly FAIR (Figure 1).

Historically, a lot of this metadata would simply be described to various degrees of thoroughness throughout a paper, but not directly linked to a particular raw or analyzed data file. While this practice might have been sufficient to interpret one small dataset at a time, it no longer serves biology well today, and adds uncertainty where it need not exist. That is especially true when combining different datatypes within a study (vertical or multi-modal data integration; Figure 2) or across studies (horizontal integration), where the statistical uncertainty associated with data from tissues from the same animal would be different from that from different animals or studies even if the animals were treated similarly.

Open-source data management systems can be found in NextSEEK (32) which allows public metadata sharing on FAIRDOMHub (33), or the Open Science Framework (34), though the latter is currently more suited for the social sciences. A similar, but more powerful alternative is Fairspace (35) provided commercially by The Hyve which supports the cancer research community's c-Bioportal (36, 37). In principle, metadata collected as part of standard electronic notebook keeping can be easily exported when and where needed, facilitating FAIR data sharing through any of these systems.



UID	D.SEQ-1-1
FileName	SRR22257136
NumReads	20891247688
Sequencer	Illumina NovaSeq6000
Parent	DNA-1-1

UID	DNA-1-1
Name	TIS1-1_DNA1_R1.fastq.gz
Type	Illumina Library
Protocol	P.SHA-220819-V1_SeqWellHughes.pdf
Parent	TIS-1-1

UID	TIS-1-1
Name	NHP-1-190507-LUL GM
Organ	Lung
Organ Detail	Left Lower Lobe
Parent	NHP-1

UID	NHP-1
Name	NHP-1
Species	Macaca fascicularis
Cohort	Untreated
Sex	Male

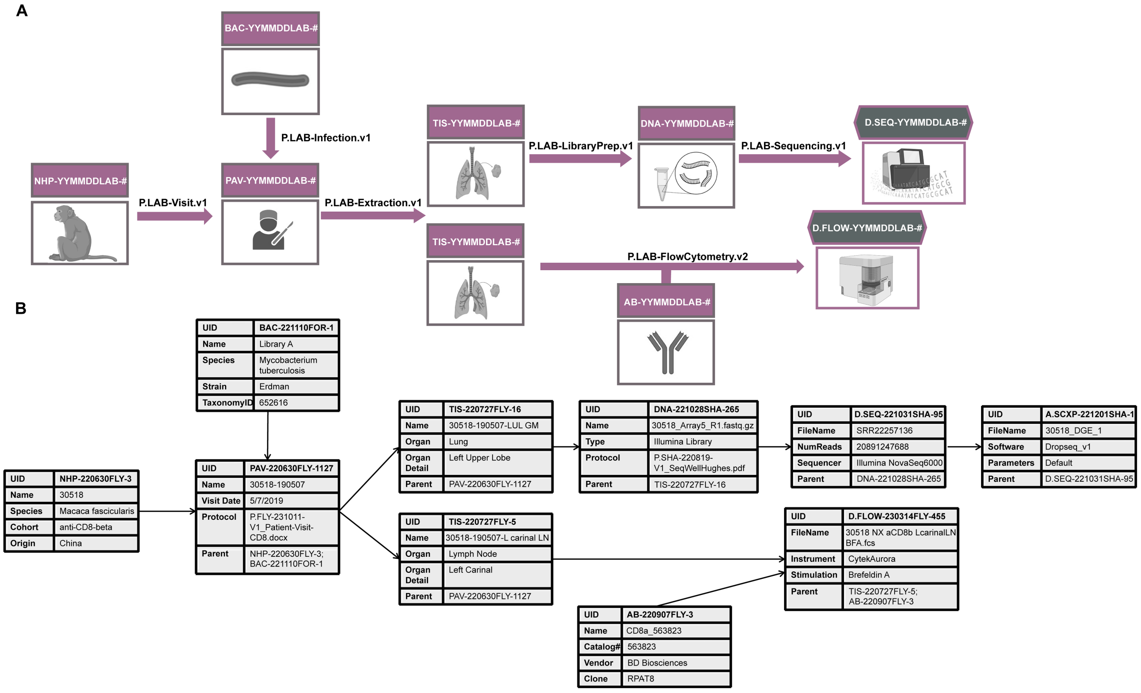
UID	D.SEQ-2-1-1
FileName	SRR17932971644
NumReads	20891247688
Sequencer	Illumina NovaSeq6000
Parent	DNA-111-1

UID	DNA-2-1
Name	TIS111-1_DNA1_R1.fastq.gz
Type	Illumina Library
Protocol	P.SHA-220819-V1_SeqWellHughes.pdf
Parent	TIS-111-1

UID	TIS-2-1
Name	30518-190507-LUL GM
Organ	Lung
Organ Detail	Left Lower Lobe
Parent	NHP-111

UID	NHP-2
Name	30518
Species	Macaca fascicularis
Cohort	Mtb infected
Sex	Female

FIGURE 1
Relational databases help link metadata from multi-step experiments. Relational databases are similar to Russian dolls, nested in a particular order. They allow researchers to capture metadata at every experimental step. For example, linking sequencing data to the cDNA library, tissue (lung biopsy) and animal (non-human primate, NHP) from which it came from. The metadata for the two different sequencing files helps future users realize that the main difference between the sequenced samples is that they come from different animals that vary by sex and treatments, despite being from the same species.



UID	BAC-221110FOR-1
Name	Library A
Species	Mycobacterium tuberculosis
Strain	Erdman
TaxonomyID	652616

UID	NHP-220630FLY-3
Name	30518
Species	Macaca fascicularis
Cohort	anti-CD8-beta
Origin	China

UID	PAV-220630FLY-1127
Name	30518-190507
Visit Date	5/7/2019
Protocol	P.FLY-231011-V1_Patient-Visit-CD8.docx
Parent	NHP-220630FLY-3; BAC-221110FOR-1

UID	TIS-220727FLY-16
Name	30518-190507-LUL GM
Organ	Lung
Organ Detail	Left Upper Lobe
Parent	PAV-220630FLY-1127

UID	TIS-220727FLY-5
Name	30518-190507-L carinal LN
Organ	Lymph Node
Organ Detail	Left Carinal
Parent	PAV-220630FLY-1127

UID	DNA-221028SHA-265
Name	30518_Array5_R1.fastq.gz
Type	Illumina Library
Protocol	P.SHA-220819-V1_SeqWellHughes.pdf
Parent	TIS-220727FLY-16

UID	AB-220907FLY-3
Name	CD8a_563823
Catalog#	563823
Vendor	BD Biosciences
Clone	RPAT8

UID	D.SEQ-221031SHA-95
FileName	SRR22257136
NumReads	20891247688
Sequencer	Illumina NovaSeq6000
Parent	DNA-221028SHA-265

UID	A.SCXP-221201SHA-1
FileName	30518_DGE_1
Software	Dropseq_v1
Parameters	Default
Parent	D.SEQ-221031SHA-95

UID	D.FLOW-230314FLY-455
FileName	30518 NX aCD8b LcarinalLN BFA.fcs
Instrument	CytekAurora
Stimulation	Brefeldin A
Parent	TIS-220727FLY-5; AB-220907FLY-3

FIGURE 2
An example of how relational databases can be used to collect metadata for multi-modal data generation. **(A)** An experiment in which a lung biopsy (TIS) from a non-human primate (NHP) infected with Mycobacterium tuberculosis (BAC) is sequenced (D.SEQ) and analyzed by flow cytometry (D.Flow). DNA refers to the cDNA library sent for sequencing, AB refers to the antibody used in the flow cytometry analysis. The experimental protocol describing how each step was conducted is captured in a file denoted with a "P." suffix and referenced in the Protocol metadata field. **(B)** An example of some metadata fields to be collected in association with each step of the research process. These fields are by no means comprehensive. Find more detailed metadata fields for a similar experiment at https://fairdomhub.org/studies/1134.

Standardization enables interoperability

Data integration has long been the focus of computational biologists, to various degrees of success (38–40) in part -some argue- because of a combination of poor data quality, experimental design and metadata availability (41–43). With the continuous increase in integratable data modalities of relevance to systems immunology, it is more important than ever to standardize how metadata is collected within and across experiments, and harmonize the vocabularies used for annotation. Unfortunately, this is no easy task. Agreeing on metadata standards for a new data type or experimental format often involves hours of discussion between experimental and computational biologists and should happen before data collection begins to avoid discrepancies down the line. This is especially true in the case of nascent technologies and requires an ability to foresee potential uses beyond what the data were originally planned for. Doing this on a field-wide level is even more challenging and requires strong vision and leadership.

For popular data types, specialist repositories often exist and some enforce the use of common data elements (44) (CDEs; pre-defined variables and acceptable values). However, these CDE are often different between repositories which makes vertical integration difficult and is further complicated by the fact that many siloed repositories offer no inherent way to link data from the same samples. This is also true in the case of inherently multi-modal measurements such as sequencing-based spatial transcriptomic technologies, which rely on a complementary set of sequencing and imaging data (45). In such cases, and in the absence of a repository for multi-modal data, the imaging data would be stored in one repository and the sequencing data in another, and would need to be linked through a public-facing database such as FAIRDOMHub (33) to be of future use (See Table 1 for examples of existing repositories for spatial transcriptomic data). Furthermore, the metadata collected by repositories is often too sparse for meaningful analysis as they tend to focus on capturing common points of variation across experiments and not much of the nuance. Generalist repositories, being data agnostic by design, are even less suited for standardized metadata collection. Thus, it is left to depositors to decide what they think is important metadata to share, and to future users to harmonize across datasets which can be very difficult without prior standardization or the release of good dictionaries with every dataset.

Data dictionaries clearly define each field as well as their possible values such that they are comprehensible to someone who was not part of the original study or is new to the field. It also facilitates standardization, which is why CDEs (44) are broadly useful. As an example, species could always be referred to by their NCBI Taxonomical ID and Latin name (73), their geographical origins by ISO 3166 codes (74), proteins by their Uniport ID (75), and antibodies by their company of source plus catalogue number, epitope, and conjugate. To avoid reinventing the wheel, researchers should reach for CDEs and standardized vocabularies in the public domain and share the ones they develop publicly for the benefit of others.

Curation ensures data trustworthiness

Depending on the resources available, data sharing done well can be a time and labor-intensive process involving several people, especially when using infrastructure that is not built-for-purpose. Thus, it is important for researchers to focus on sharing high-quality (well annotated and from well-designed experiments) irrespective of whether their own analysis of the data supports their original hypotheses. This should be done with an eye towards enhancing reproducibility, but also the reuse of data for mechanistic modeling, machine learning (ML) and deep learning (DL) applications, which will undoubtedly increase the impact of the data on the long-term. As mentioned above, curated data should include both negative and positive data to avoid biased training datasets that do not allow the development of models that are broadly generalizable. If time is tight, researchers should prioritize multiplexed datasets (multi-parameter flow cytometry/CyToF, high-throughput sequencing of all kinds, mass spectrometry, high volume imaging data, array-based data, cytokine panels, systems serology data to name a few), which are most useful for data hungry ML/DL applications, as well as data acquired from experiments that are difficult to reproduce without an abundance of resources or access to highly specialized infrastructure.

Code, model and parameter sharing facilitate reproducibility, interpretation, and informed reanalysis and meta analyses

Also important for reproducibility, integration, and the informed interpretation of analyzed data is capturing the complexity of data processing and computational analyses occurring post-generation of raw data. Unlike classical statistical tests familiar to many biologists [eg: the parametric and non-parametric tests pre-programmed in many available software suites such as Excel (76), Google Sheets (77), and GraphPad Prism (78)], many OMICs and most multi-OMIC analyses are far from standardized. Furthermore, compute environment, the choice of software, software version, and user-defined parameters can significantly affect the final output (1). For these reasons, researchers need to precisely document and share all aspects of a workflow including the code (including version number if using a publicly available package) and exact parameters used to analyze a particular dataset, with clear descriptions of any non-standard steps maybe as comments between blocks of code. This can be accomplished in a variety of ways but is greatly facilitated by the use of community workflows such as nf-core (79), containerized compute environments like Docker (80) and Singularity/Apptainer (81) shared in container repositories like DockerHub (82). Also useful are package and environment managers such as Bioconda (83). Jupyter (84) or Rstudio (85) notebooks shared and managed in code repositories such as Github (86) provide a method for sharing both standard and custom analyses, though this practice does not guarantee reproducibility across computing environments since Github does not enforce rigorous testing to ensure deposited packages are performant.

Researchers should also share their trained models given how time and computationally intensive this can be, in addition to the data on which they were trained to ensure full transparency and inform users' understanding of sources of bias or underperformance. Parameterized mechanistic models can be shared on BioModels (87), while their machine and deep learning equivalents can be shared and deployed on Hugging Face (88).

Resources that enable good FAIR data stewardship

Infrastructure

To support FAIR data practices institutions must facilitate accurate data and metadata collection with little time and effort on researchers' side. Research institutions and funders also need to account for the increasing specialization that necessitates collaboration between labs. To enable that, there is a need for a radical change in infrastructure to support an evolution to the "decentralized digital Lab with a human in the loop". In this model, laboratory infrastructure is set up such that data acquisition and import is largely automated within and between labs with the proper agreements in place, meaning scientists spend less time generating and managing data and more time curating and analyzing it. The first step towards that has been a slow-to-start but accelerating shift from paper to electronic lab notebooks (ELNs), catalyzed by the evolution of user-friendly digital platforms such as Benchling (89). Benchling's cloud-based ELN system now allows independent users anywhere in the world to share experimental templates, as well as track reagents, samples, and data through a shared registry, while linking these features through a set of relational databases ensuring data provenance is continuous and available to all who have access. Add to that the addition of features that allow the integration of lab instruments such that the data coming off them can be directly stored in the cloud, and the effort of moving data, linking to metadata, and -eventually- sharing no longer seems as daunting. Other providers such as L7 informatics are catching up (90). With more players in the Digital Lab eco-system the future of FAIR data sharing is looking promising (Figure 3). Quite importantly, this also facilitates data interpretation and saves research teams hours of lengthy discussion about how data was generated and handled.

Dedicated cloud computing platforms for biology are also emerging to compliment the shift to data-intense, decentralized and collaborative life science research, including Cirro (of the Fred Hutchinson Institute) (91), DNANexus (a techbio start up) (92), LatchBio (a techbio start up) (93), Terra (of the Broad Institute) (94), and L7 Informatics (also a techbio start up) (90). These platforms allow researchers to run complex analysis workflows in the cloud but in a more user-friendly environment than what's offered directly by cloud providers and are customized to biologists' needs. With the data already in the cloud, running such analyses is now possible without the need to duplicate and shuffle around large volumes of data between collaborators, and -in principle- facilitates analyses that respect institutional and national data governance requirements. Cloud providers vow they take data security seriously, and the likes of the NHS, FDA and NIH are beginning to trust them with data for

TABLE 1 Key data repositories useful for systems immunology.

Data Type	Repository	Key Features	Number of datasets as of May 2024	Date of establishment
Mass spectrometry (MS)-based Proteomics	PRIDE: PRoteomics IDEntifications Database (46) (https://www.ebi.ac.uk/pride/)	Direct submission allowed, data visualization and annotation tools.	26847	2005
	MassIVE (47) (https://massive.ucsd.edu/)	Direct submission allowed, data analysis tools.	15,231	N/A
	PeptideAtlas (48) (https://peptideatlas.org/)	Curated database, no data analysis tools.	N/A	2006
	Panorma (49) (https://panoramaweb.org/)	Data from targeted proteomics experiments, direct submissions allowed, tools for designing and analyzing targeted proteomics experiments.	596	2014
	iProX (50) (https://www.iprox.cn/)	Direct submission allowed, no data analysis tools.	4792 Projects (3602 Public Projects)	2019
	JPOST (51) (https://repository.jpostdb.org/)	Direct submission allowed, no data analysis tools.	2671 projects	2017
MS-based Metabolomics	MetaboLights (52) (https://www.ebi.ac.uk/metabolights)	Direct submission allowed, no data analysis tools	1496	2012
	National Metabolomics Data Repository (NMDR; https://www.metabolomicsworkbench.org/data/DRCCDataDeposit.php)	Direct submission allowed, no data analysis tools.	2788	2020
ELISA, ELISPOT, Luminex	ImmPort (53) (https://www.immport.org/)	Immunology-focused, direct submission allowed, rich metadata in relational database, no data analysis tools.	262, 54, 61	2018
Flow Cytometry	ImmPort (53) (https://www.immport.org/)	Immunology-focused, direct submission allowed, rich metadata in relational database, no data analysis tools.	257	2018
	FlowRepository (54) (https://flowrepository.org/)	Direct submission allowed, follows MIFlowCyt standard, endorsed by International Society for Advancement of Cytometry (ISAC), no data analysis tools.	~2125	2012
Imaging	Image Data Resource (55) (IDR; https://idr.openmicroscopy.org/)	Direct submission allowed, handles variety of image types, no data analysis tools.	127 Studies	2017
	The Cell (CIL-CCDB) (56): (http://www.cellimagelibrary.org/)	Curated database, no data analysis.	57	2012
	Cancer Imaging Archive (TCIA) (57): (https://www.cancerimagingarchive.net/)	Data de-identified, allows direct submissions, no analysis tools.	N/A	2013
NGS and array data	Sequence Read Archive (58) (SRA; https://www.ncbi.nlm.nih.gov/sra)	Allows direct submissions of sequencing data, no analysis tools.	N/A	2007
	Database of Genotypes and Phenotypes (59) (dbGAP; https://www.ncbi.nlm.nih.gov/gap/)	Allows direct submissions of sequencing data, controlled access repository for human genotype/phenotype data.	309 general use studies ie: sharable according to these (60) terms and nothing else.	2006
	The Bioinformation and DNA Data Bank of Japan (61) (DDBJ; https://www.ddbj.nig.ac.jp/)	Allows direct submissions of sequencing and array data, provides advanced search functionalities and built-in analysis tools.	4,250,864,039 Sequences	1987
	European Nucleotide Archive (62) (ENA; (https://www.ebi.ac.uk/ena)	Allows direct submission of sequencing and data, no data analysis tools.	4.6 billion Sequences	1982
	Gene Expression Omnibus (63) (GEO; https://www.ncbi.nlm.nih.gov/geo/)	Allow direct submissions of sequencing and MIAME-compliant array data as well	4348	2000

(Continued)

TABLE 1 Continued

Data Type	Repository	Key Features	Number of datasets as of May 2024	Date of establishment
		as processed data, some data analysis tools.		
Single Cell Sequencing	Single Cell Portal: (https://singlecell.broadinstitute.org/)	Allows submission of sequencing and processed single cell data files, data visualization and analysis tools.	670 total studies found	2018
	Single Cell Expression Atlas (64) (https://www.ebi.ac.uk/gxa/sc/home)	Curated database, data visualization and analysis.	355	2018
Spatial Transcriptomics	CROST (65) (https://ngdc.cncb.ac.cn/crost/home)	Curated database, supports different technologies, rich suite of data analysis and visualization tools.	182	2024
	Spatial DB (66) (http://www.spatialomics.org/SpatialDB/)	Curated database, supports different technologies, some data analysis tools.	24	
	STOmicsDB (67) (https://db.cngb.org/stomics/)	Curated database, allows direct submission, some data visualization and analysis tools.	228	
	Spatial Omics DataBase (68) (SODB; https://gene.ai.tencent.com/SpatialOmics/)	Curated database, supports different technologies, some data visualization and analysis tools.	3145	2023
	Aquila (69) (https://aquila.cheunglab.org)	Curated database, allows direct submission, some data visualization and analysis tools.	110	2023
Single Cell Sequencing	Single Cell Portal (70) (https://singlecell.broadinstitute.org/)	Allows submission of sequencing-based spatial transcriptomic data, data visualization and analysis tools.	670 total studies found	2018
Multi-modal OMICs	Single Cell Atlas (71) (https://www.singlecellatlas.org/)	Curated database, multiple data types, data visualization and analysis.	NA	2024
Generalist	Zenodo – commercial (https://zenodo.org/)	50GB dataset limit, any file type, GitHub integration, DOI creation, version control, immediate release, usage statistics.	1,609 Projects	2013
	Figshare -commercial: (https://figshare.com/)	20 GB per user, any file type, DOI creation, version control, private and public release, usage statistics.	N/A	2012
	BioStudies (72) (https://www.ebi.ac.uk/biostudies/)	Allows the integration of metadata, orphan data, and data found in other EBI databases and link to a paper.	2,398,047	2015
	FAIRDOMHub (33) (https://fairdomhub.org)	Allows the integration of metadata, orphan data, and data found in other databases and link to a paper.	402 projects	2017

megaprojects such as the UK Biobank (95) and PrecisionFDA (96) who use DNAnexus and the NIH's All of Us Research program (97) that uses Terra (Figure 3). Enticingly for the computational biologists, these platforms also provide impressive compute that scales to increasingly large and complex models, come with customizable and pre-installed pipelines that save researchers hours of set-up time, and automate log generation which allows tracking the analyses done on every dataset together with the parameters used making it easy to trace how results were derived. In addition, some of these platforms support the integration of Jupyter notebooks which, as mentioned above, allow users to run and share their own custom code within those environments and share them when needed.

Less attractive is the price for cloud storage and compute which becomes an ongoing expense liable to immense runaway costs (98), especially at the hands of less experienced users who are the majority at academic institutions today. These costs could be overcome -in part- by better training, negotiating university/funder-wide contracts with cloud platforms, and could be off-set by long-term savings in personnel, maintenance, and upgrades. However, this does leave academics at the mercy of tech oligarchs such as Amazon [providers of AWS (99)], Alphabet [providers of Google Cloud (100)], and Microsoft [providers of Azure (101)]. At the moment, it is also unclear how easy migration between any of these platforms will be if they fail to meet future user needs. That said, competition in the

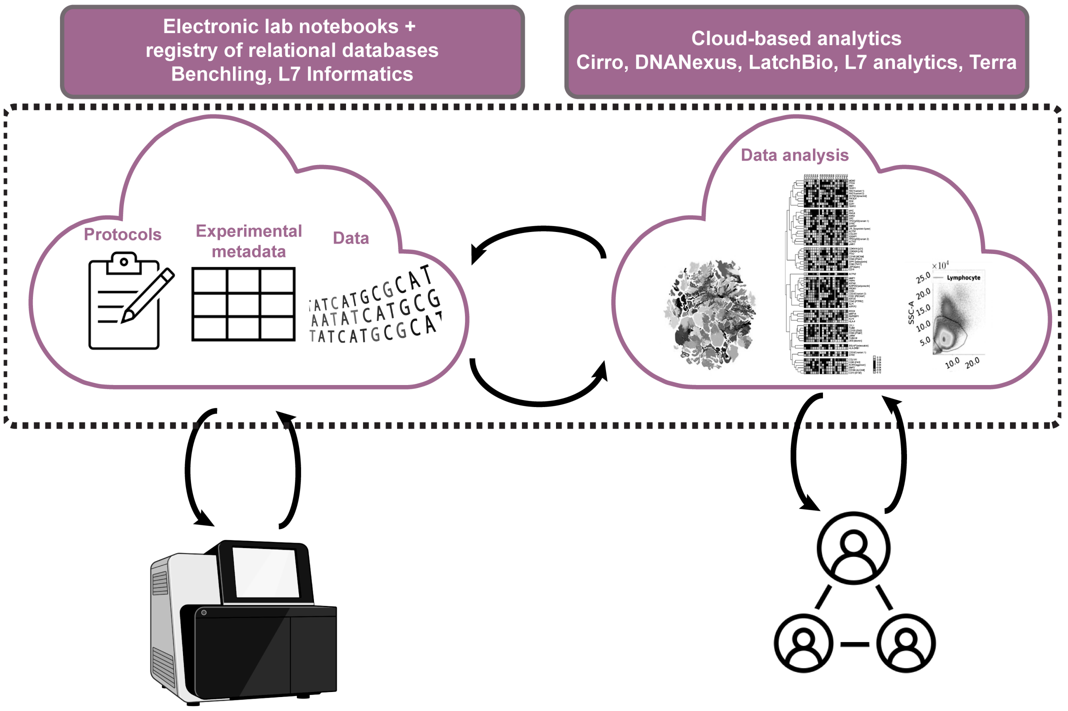

Electronic lab notebooks +
registry of relational databases
Benchling, L7 Informatics
Cloud-based analytics
Cirro, DNANexus, LatchBio, L7 analytics, Terra
Protocols
Experimental metadata
Data
ATCATGCGCAT
AATATCATGCG
TATCATGCGCAT
Data analysis
Lymphocyte
SSC-A
×10⁴

FIGURE 3
Overview of cloud-based infrastructure for digital labs with humans in the loop. Cloud-based electronic documentation systems with registries for animals, samples, reagents, equipment allow facile linking within lab notebooks. This makes experimental protocols transparent and facilitates FAIR metadata and data collection and sharing. Data from lab equipment is imported directly into the registry and can also be linked in lab notebooks saving time and effort, while minimizing human error. Metadata and data can be pushed to cloud computing platforms that allow collaborative and transparent data analysis.

infrastructure-as-a-service space is increasing because of ubiquitous demand across a variety of industries which will hopefully spur technological innovation and push prices down, democratizing access to infrastructure-as-a-service in the long-term.

The development of equally powerful open-source alternatives, continuously developed by and for the research community would be ideal. Unfortunately, funding such efforts is costly requiring a hefty upfront investment from governments or philanthropists and would take years adding more distance between them and their well-developed commercial counterparts. Once developed, long-term sustainability could be possible by licensing that allows free academic/non-profit usage and paid licensing in the case of for-profit entities similar to the Rosetta Commons approach (102).

Personnel

With the emergence of infrastructure-as-a-service, better ELNs and digital lab management software (also referred to as LIMS), data, compute, and metadata are all now connectable and shareable with relative ease. But the transition to this new model is no easy feat, mostly because of the need for complex and often continuous change management since academic research inherently involves training inexperienced individuals and high turnover.

To facilitate this, the first kind of position universities need to create is that of the Data Officer. This individual outlines university-wide policy, and ensures it aligns with national and international mandates and legal frameworks. Together with the IT officer, they can help coordinate the roll out and adoption of the needed technology for the shift to Digital Labs across the university while vetting different vendors and platforms - all in coordination with Department Data Managers. The latter help roll out these changes within their departments and communicate the importance of FAIR data sharing to researchers of different disciplines, as well as work with them to develop and share best-practices that make FAIR data sharing a natural part of researchers' workflow with the help of enabling infrastructure. They also provide guidance regarding the choice of private storage and public data repositories depending on the types of data generated (eg: sequencing, flow cytometry, imaging), its sensitivity (eg: clinical versus pre-clinical data), and its stage in the research life cycle (pre- *vs* post-publication). On the other hand, the university/department IT officers works on optimizing on premise compute and data storage requirements to adapt to a shift to the cloud. For larger, data-intensive departments, it might be necessary to have lab data managers who work even more closely with the researchers on the day-to-day. That said, buy-in from PIs is absolutely necessary for the success of these efforts and communicating the importance of good data management practices early on in every project is immensely important for labs' long-term success. This is particularly important in academic contexts, where lab turnover is high, and data often pass multiple hands before it ends up in a paper or in the public domain.

Common types of data and their repositories

For systems immunologists in particular and life science researchers in general, some of the most important data types today include multiplexed flow cytometry and CyTOF, Luminex, systems serology [Luminex-based antibody profiling assays (103, 104)], single cell RNASeq, bulk RNASeq data, and imaging data, for which specialized repositories currently exist. For more nascent fields or in the case of orphan datatypes researchers are encouraged to deposit their datasets in generalist repositories, until a dedicated repository is developed (Table 1).

Considerations for consortia

While the guidelines outlined above are broadly generalizable, they are highly relevant for interdisciplinary academic consortia which are somewhat of a special case for three main reasons: (1) data sharing between labs in almost real-time (as opposed to at the time of publication) is important for consortia to achieve their goals as experimental and biological versus mathematical and computational expertise tend to be distributed, (2) physical samples are often exchanged between labs so tracking sample as well as data provenance at scale is key to data integrity, (3) timely exchange of knowledge requires close communication across

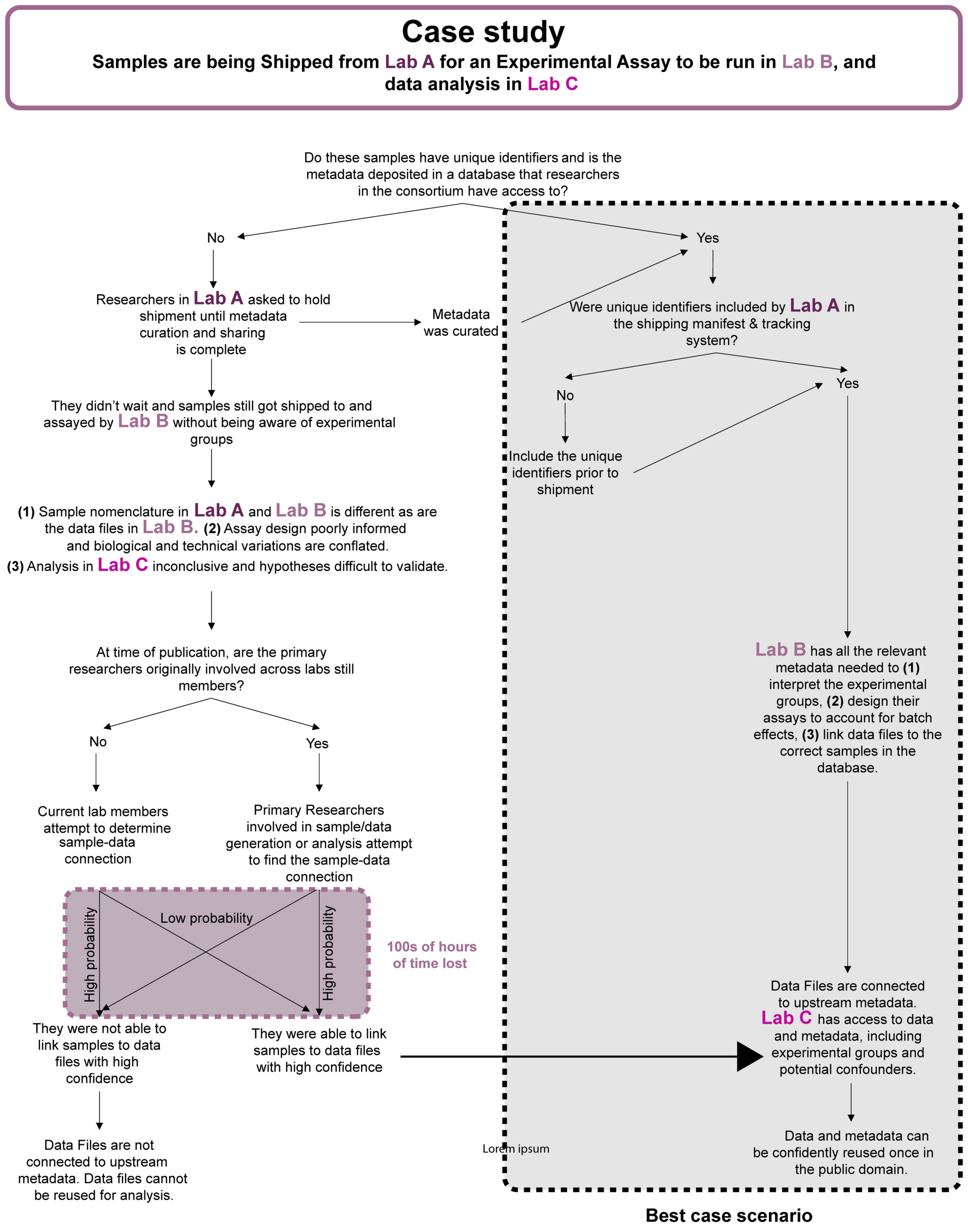

Case study
Samples are being Shipped from Lab A for an Experimental Assay to be run in Lab B, and data analysis in Lab C
Do these samples have unique identifiers and is the metadata deposited in a database that researchers in the consortium have access to?
No
Yes
Researchers in Lab A asked to hold shipment until metadata curation and sharing is complete
Metadata was curated
Were unique identifiers included by Lab A in the shipping manifest & tracking system?
No
Yes
They didn't wait and samples still got shipped to and assayed by Lab B without being aware of experimental groups
Include the unique identifiers prior to shipment
(1) Sample nomenclature in Lab A and Lab B is different as are the data files in Lab B. (2) Assay design poorly informed and biological and technical variations are conflated.
(3) Analysis in Lab C inconclusive and hypotheses difficult to validate.
At time of publication, are the primary researchers originally involved across labs still members?
No
Yes
Lab B has all the relevant metadata needed to (1) interpret the experimental groups, (2) design their assays to account for batch effects, (3) link data files to the correct samples in the database.
Current lab members attempt to determine sample-data connection
Primary Researchers involved in sample/data generation or analysis attempt to find the sample-data connection
Low probability
High probability
High probability
100s of hours of time lost
They were not able to link samples to data files with high confidence
They were able to link samples to data files with high confidence
Data Files are connected to upstream metadata. Lab C has access to data and metadata, including experimental groups and potential confounders.
Data Files are not connected to upstream metadata. Data files cannot be reused for analysis.
Lorem ipsum
Data and metadata can be confidently reused once in the public domain.
Best case scenario

FIGURE 4
Workflow diagram describing how suboptimal practices can affect data integrity, analytical rigor, and waste valuable research time. Suboptimal practices also compromise FAIR data sharing at the end of a study.

disciplines to move project goals forward and course-correct as needed.

To address the first and second point, setting up a unified or at least interoperable, digital infrastructure at the onset is key as it allows members of the consortium to share lab notebooks, reagents, and data. Harmonization and standardization becomes easier to achieve and enforce, and analyses can also be shared. This allows for internal transparency, facilitates collaboration, troubleshooting, and corrections, and ultimately multi-modal data analysis. Eventually, sharing the data and knowledge in the public domain also becomes easier (Figure 4).

Assigning a data officer and data base manager for the consortium is key. Together they need to establish a system that allows centralized sample and data tracking and work with each lab's data manager and/or individual researchers to collate curated data and ensure the accompanying metadata is accurate, standardized, and complete. In instances when access to unified infrastructure is prohibitively expensive, individual components can be strung together. For example, they could set up a shared Dropbox or Google Drive account where all the consortium's curated data is collected until it can be shared in the appropriate repositories. Setting up a local instance of NextSEEK would facilitate metadata collection in a relational database, after researchers fill out easy to use spreadsheet-based templates. Tracking samples shipped between sites can be done using tools such as Qualtrics (105) or Google Forms (106). To facilitate early and accurate collection of metadata about samples, only samples for which unique identifiers and a comprehensive set of metadata has been collected should be shipped to other sites. Data should only be shared when all the metadata is complete, they are uploaded to a repository (privately) or a common drive, and linked in the central database. Recipients can then use the unique identifiers to look up key information about each sample/dataset in the database and find all the necessary metadata. Access to the data before it is public can be decided as needed since considerations may vary depending on project needs or data type and source, eg: human versus non-human sequencing data.

To address the third point, establishing recurrent meetings that bringing together researchers of complementary expertise to discuss experimental design, data analysis, next steps, and synthesize information is important. This helps ensure that the data is analyzed and interpreted in a meaningful way, as well as used to inform the design of appropriate follow-up experiments.

FAIR data sharing between interdisciplinary teams is critical to the responsible development and deployment of AI

To summarize, the life sciences are on the cusp of a transformation to a data-intense field that requires experimental and computational biologists to work together and make sense of large swaths of data using mechanistic, ML and DL models. This will allow researchers to generate new insights that drive biomedical research forward in ways and at a scale previous not possible. Enabling this involves embracing a collaborative and digital-first mentality to sustain the development of data hungry, unbiased, generalizable models that are helpful to biomedical researchers. We argue that the future of such a transformation involves a shift to the Digital Lab and decentralized FAIR data sharing and compute to enable broader collaboration across disciplines. Despite the complexity of the feat, it is necessary to ensure that data is scrutinized, used, and re-used to the best extent possible, maximizing return on investment in the research enterprise for the benefit of all.

Author contributions

DM: Conceptualization, Data curation, Investigation, Project administration, Supervision, Visualization, Writing – original draft, Writing – review & editing. JL: Investigation, Writing – review & editing. CD: Visualization, Writing – review & editing. NE: Data curation, Investigation, Writing – review & editing. CW: Writing – review & editing. MG: Writing – review & editing. DL: Supervision, Writing – review & editing. SF: Conceptualization, Supervision, Writing – review & editing. SL: Conceptualization, Supervision, Writing – review & editing, Writing – original draft.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. All authors are funded in whole or in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. 75N93019C00071.

Acknowledgments

We would like to thank the teams at Benchling, Cirro, LatchBio, and Terra for insightful discussions and clarifications relating to their platforms.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Wilkinson MD, Dumontier M, Aalbersberg J, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. (2016) 3:160018. doi: 10.1038/sdata.2016.18

2. Tiku N, Schaul K, Chen SY. *These fake images reveal how AI amplifies our worst stereotypes. Washington Post*. Available online at: https://www.washingtonpost.com/technology/interactive/2023/ai-generated-images-bias-racism-sexism-stereotypes/ (Accessed May 20, 2024).

3. Metz C. OpenAI unveils new chatGPT that listens, looks and talks. *New York Times*. (2024).

4. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. (2021) 596:583–9. doi: 10.1038/s41586-021-03819-2

5. Jumper J, Hassabis D. The protein structure prediction revolution and its implications for medicine: 2023 albert lasker basic medical research award. *JAMA*. (2023) 330:1425–6. doi: 10.1001/jama.2023.17095

6. *All Nobel Prizes 2024*. NobelPrize.org. Available online at: https://www.nobelprize.org/all-nobel-prizes-2024/ (Accessed December 13, 2024).

7. *Index of /pub/databases/uniprot/previous_releases/release-2020_01/uniref*. Available online at: https://ftp.ebi.ac.uk/pub/databases/uniprot/previous_releases/release-2020_01/uniref/ (Accessed May 2, 2024).

8. Steinegger S, Johannes. BFD. Available online at: https://bfd.mmseqs.com/ (Accessed May 2, 2024).

9. Mirdita M, von den Driesch L, Galiez C, Martin MJ, S ding J, Steinegger M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res*. (2017) 45:D170–6. doi: 10.1093/nar/gkw1081

10. Richardson L, Allen B, Baldi G, Beracochea M, Bileschi ML, Burdett T, et al. MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res*. (2023) 51:D753–9. doi: 10.1093/nar/gkac1080

11. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. *Nucleic Acids Res*. (2000) 28:235–42. doi: 10.1093/nar/28.1.235

12. Chowdhury R, Bouatta N, Biswas S, Floristean C, Kharkar A, Roy K, et al. Single-sequence protein structure prediction using language models and deep learning. *Nat Biotechnol*. (2022) 40:1617–23. doi: 10.1038/s41587-022-01432-w

13. Polonsky K, Pupko T, Freund NT. Evaluation of the ability of alphaFold to predict the three-dimensional structures of antibodies and epitopes. *J Immunol*. (2023) 211:1578–88. doi: 10.4049/jimmunol.2300150

14. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res*. (2022) 50:D439–44. doi: 10.1093/nar/gkab1061

15. Google Deep Mind. (2024). Available online at: https://deepmind.google/technologies/alphafold/ (Accessed January 3, 2024).

16. Varadi M, Velankar S. The impact of AlphaFold Protein Structure Database on the fields of life sciences. *Proteomics*. (2023) 23:e2200128. doi: 10.1002/pmic.202200128

17. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*. (2018) 34:4121–3. doi: 10.1093/bioinformatics/bty407

18. Chen L. *What Works to Control COVID-19? Econometric Analysis of a Cross-Country Panel*. (2020).

19. Scottish Government. Coronavirus (COVID-19) Scotland's Strategic Framework update – February 2022: evidence paper. (2022). Available online at: https://www.gov.scot/publications/evidence-paper-accompany-coronavirus-covid-19-scotlands-strategic-framework-update-february-2022/ (Accessed May 8, 2024).

20. World Health Organization. *Genomic Sequencing of SARS-CoV-2: A Guide to Implementation for Maximum Impact on Public Health, 8 January 2021*. World Health Organization (2021). Available online at: https://www.who.int/publications/i/item/9789240018440 (Accessed May 2, 2024).

21. Karmarkar EN, Blanco I, Amornkul PN, DuBois A, Deng X, Moonan PK, et al. Timely intervention and control of a novel coronavirus (COVID-19) outbreak at a large skilled nursing facility—San Francisco, California, 2020. *Infect Control Hosp. Epidemiol*. (2021) 42:1173–80. doi: 10.1017/ice.2020.1375

22. Bennett H, Stewart-Jones G, Narayanan E, Carfi A, Metkar M, Presnyak V, et al. *Coronavirus RNA vaccines and methods of use. World Intellectual Property Organization*. Cambridge University Press. (2020). WO2021159130A2.

23. *doi Handbook*. Available online at: https://www.doi.org/doi-handbook/HTML/index.html (Accessed May 8, 2024).

24. Torralba A, Efros AA. (2011). Unbiased look at dataset bias, in: *CVPR 2011*, IEEE. pp. 1521–8. doi: 10.1109/CVPR.2011.5995347

25. *NOT-OD-21-013: Final NIH Policy for Data Management and Sharing*. Available online at: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html (Accessed May 9, 2024).

26. National Science Foundation. Preparing Your Data Management Plan - Funding at NSF. (2024). Available online at: https://new.nsf.gov/funding/data-management-plan (Accessed May 8, 2024).

27. The White House. *OSTP Issues Guidance to Make Federally Funded Research Freely Available Without Delay | OSTP*. The White House (2022). Available online at: https://www.whitehouse.gov/ostp/news-updates/2022/08/25/ostp-issues-guidance-to-make-federally-funded-research-freely-available-without-delay/ (Accessed May 8, 2024).

28. Bill & Melinda Gates Foundation. *Data Sharing Requirements*. Gates Open Access Policy. Available online at: https://openaccess.gatesfoundation.org/how-to-comply/data-sharing-requirements/ (Accessed May 8, 2024).

29. Chan Zuckerberg Initiative. *Funding Scientific Research and Building Transformative Technologies*. Chan Zuckerberg Initiative. (2024). Available online at: https://chanzuckerberg.com/science/our-values-approach/.

30. Wellcome. Data, software and materials management and sharing policy - Grant Funding. (2025). Available online at: https://wellcome.org/grant-funding/guidance/policies-grant-conditions/data-software-materials-management-and-sharing-policy (Accessed January 3, 2025).

31. European Comission. Open Data, Software and Code Guidelines. Open Research Europe. Available online at: https://open-research-europe.ec.europa.eu/for-authors/data-guidelines (Accessed May 9, 2024).

32. Pradhan D, Ding H, Zhu J, Engelward BP, Levine SS. NExtSEEK: extending SEEK for active management of interoperable metadata. *J Biomolecular Techniques*. (2022) 33. doi: 10.7171/3fc1f5fe.db404124

33. Wolstencroft K, Krebs O, Snoep JL, Stanford NJ, Bacall F, Golebiewski MG, et al. FAIRDOMHub: a repository and collaboration environment for sharing systems biology research. *Nucleic Acids Res*. (2017) 45:D404–7. doi: 10.1093/nar/gkw1032

34. Foster ED, Deardorff A. Open science framework (OSF). *J Med Libr Assoc*. (2017) 105:203–6. doi: 10.5195/jmla.2017.88

35. *The Hyve*. Available online at: https://thehyve.nl/services (Accessed May 9, 2024).

36. *cBioPortal for Cancer Genomics*. Available online at: https://www.cbioportal.org/ (Accessed May 9, 2024).

37. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discovery*. (2012) 2:401–4. doi: 10.1158/2159-8290.CD-12-0095

38. Izzo F, Myers RM, Ganesan S, Mekerishvili L, Kottapalli S, Prieto T, et al. Mapping genotypes to chromatin accessibility profiles in single cells. *Nature*. (2024) 629:1149–57. doi: 10.1038/s41586-024-07388-y

39. Lance C, Luecken MD, Burkhardt DB, Cannoodt R, Rautenstrauch P, Laddach A, et al. (2022). Multimodal single cell data integration challenge: Results and lessons learned, in: *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*, PMLR. pp. 162–76. Available online at: https://proceedings.mlr.press/v176/lance22a.html.

40. Boehm KM, Aherne EA, Ellenson L, Nikolovski I, Alghamdi M, Vázquez-García IV. Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. *Nat Cancer*. (2022) 3:723–33. doi: 10.1038/s43018-022-00388-9

41. Rajendran S, Pan W, Sabuncu MR, Chen Y, Zhou J, Wang F, et al. Learning across diverse biomedical data modalities and cohorts: Challenges and opportunities for innovation. *Patterns (N Y)*. (2024) 5:100913. doi: 10.1016/j.patter.2023.100913

42. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. *Brief Bioinform*. (2017) 19:1236–46. doi: 10.1093/bib/bbx044

43. Lipkova J, Chen RJ, Chen B, Lu MY, Barbieri M, Shao D, et al. Artificial intelligence for multimodal data integration in oncology. *Cancer Cell*. (2022) 40:1095–110. doi: 10.1016/j.ccell.2022.09.012

44. *NIH Common Data Elements (CDE) Repository*. Available online at: https://cde.nlm.nih.gov/home (Accessed May 17, 2024).

45. Williams CG, Lee HJ, Asatsuma T, Vento-Tormo R, Haque A. An introduction to spatial transcriptomics for biomedical research. *Genome Med*. (2022) 14:68. doi: 10.1186/s13073-022-01075-1

46. Perez-Riverol Y, Bai J, Bandla C, Garc a-Seisdedos D, Hewapathirana S, Kamatchinathan Selvakumar, et al. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res*. (2022) 50:D543–52. doi: 10.1093/nar/gkab1038

47. Choi M, Carver J, Chiva C, Tzouros M, Huang T, Tsai Tsung-Heng, et al. MassIVE.quant: a community resource of quantitative mass spectrometry–based proteomics datasets. *Nat Methods*. (2020) 17:981–4. doi: 10.1038/s41592-020-0955-0

48. Deutsch EW, Lam H, Aebersold R. PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep*. (2008) 9:429–34. doi: 10.1038/embor.2008.56

49. Sharma V, Eckels J, Taylor GK, Shulman NJ, Stergachis AB, Joyner SA, et al. Panorama: A targeted proteomics knowledge base. *J Proteome Res*. (2014) 13:4205–10. doi: 10.1021/pr5006636

50. Chen T, Ma J, Liu Y, Chen Zhiguang, Xiao N, Lu Yutong, et al. iProX in 2021: connecting proteomics data sharing with big data. *Nucleic Acids Res.* (2022) 50:D1522–7. doi: 10.1093/nar/gkab1081

51. Moriya Y, Kawano S, Okuda Shujiro, Watanabe Y, Matsumoto M, Takami T, et al. The jPOST environment: an integrated proteomics data repository and database. *Nucleic Acids Res.* (2019) 47:D1218–24. doi: 10.1093/nar/gky899

52. Yurekten O, Payne T, Tejera N, Amaladoss FX, Martin C, Williams M, et al. MetaboLights: open data repository for metabolomics. *Nucleic Acids Res.* (2024) 52: D640–6. doi: 10.1093/nar/gkad1045

53. Bhattacharya S, Dunn P, Thomas CG, Smith B, Schaefer H, Chen J, et al. ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. *Sci Data.* (2018) 5:180015. doi: 10.1038/sdata.2018.15

54. Spidlen J, Breuer K, Rosenberg C, Kotecha N, Brinkman RR. FlowRepository: a resource of annotated flow cytometry datasets associated with peer-reviewed publications. *Cytometry A.* (2012) 81:727–31. doi: 10.1002/cyto.a.v81a.9

55. Williams E, Moore J, Li SW, Rustici G, Tarkowska A, Chessel A, et al. Image Data Resource: a bioimage data integration and publication platform. *Nat Methods.* (2017) 14:775–81. doi: 10.1038/nmeth.4326

56. Orloff DN, Iwasa JH, Martone ME, Ellisman MH, Kane CM. The cell: an image library-CCDB: a curated repository of microscopy data. *Nucleic Acids Res.* (2013) 41: D1241–1250. doi: 10.1093/nar/gks1257

57. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The cancer imaging archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging.* (2013) 26:1045–57. doi: 10.1007/s10278-013-9622-7

58. Leinonen R, Sugawara H, Shumway M. The sequence read archive. *Nucleic Acids Res.* (2011) 39:D19–21. doi: 10.1093/nar/gkq1019

59. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, et al. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet.* (2007) 39:1181–6. doi: 10.1038/ng1007-1181

60. NIH. *DBGAP data use certification agreement.* dbgap.ncbi.nlm.nih.gov. Available online at: https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=DUC&view_pdf&stacc=phs000688.v1.p1.

61. Ara T, Kodama Y, Tokimatsu T, Fukuda A, Kosuge T, Mashima J, et al. DDBJ update in 2023: the MetaboBank for metabolomics data and associated metadata. *Nucleic Acids Res.* (2023) 52:D67–71. doi: 10.1093/nar/gkad1046

62. Burgin J, Ahamed A, Cummins C, Devraj R, Gueye K, Gupta D, et al. The european nucleotide archive in 2022. *Nucleic Acids Res.* (2022) 51:D121–5. doi: 10.1093/nar/gkac1051

63. National Center for Biotechnology Information- NCBI. NCBI GEO: archive for functional genomics data sets–update. Available online at: https://pubmed.ncbi.nlm.nih.gov/23193258/ (Accessed May 10, 2024).

64. Papatheodorou I, Moreno P, Manning J, Fuentes AM-P, George N, Fexova S, et al. Expression Atlas update: from tissues to single cells. *Nucleic Acids Res.* (2020) 48: D77–83. doi: 10.1093/nar/gkz947

65. Wang G, Wu S, Xiong Z, Qu H, Fang X, Bao Y. CROST: a comprehensive repository of spatial transcriptomics. *Nucleic Acids Res.* (2024) 52:D882–90. doi: 10.1093/nar/gkad782

66. Fan Z, Chen R, Chen X. SpatialDB: a database for spatially resolved transcriptomes. *Nucleic Acids Res.* (2020) 48:D233–7. doi: 10.1093/nar/gkz934

67. Xu Z, et al. STOmicsDB: a comprehensive database for spatial transcriptomics data sharing, analysis and visualization. *Nucleic Acids Res.* (2023) 52:D1053–61.

68. Yuan Z, Pan W, Zhao X, Zhao F, Xu Z, Li X, et al. SODB facilitates comprehensive exploration of spatial omics data. *Nat Methods.* (2023) 20:387–99. doi: 10.1038/s41592-023-01773-7

69. Zheng Y, Chen Y, Ding X, Wong KH, Cheung E. Aquila: a spatial omics database and analysis platform. *Nucleic Acids Res.* (2023) 51:D827–34. doi: 10.1093/nar/gkac874

70. Tarhan L, Bistline J, Chang J, Galloway B, Hanna E, Weitz E. Single Cell Portal: an interactive home for single-cell genomics data. *bioRxiv.* (2023). doi: 10.1101/2023.07.13.548886

71. Pan L, Parini P, Tremmel R, Loscalzo J, Lauschke VM, Maron BA, et al. Single Cell Atlas: a single-cell multi-omics human cell encyclopedia. *Genome Biol.* (2024) 25:104. doi: 10.1186/s13059-024-03246-2

72. Sarkans U, Gostev M, Athar A, Behrangi E, Melnichuk O, Ali A, et al. The BioStudies database-one stop shop for all data supporting a life sciences study. *Nucleic Acids Res.* (2018) 46:D1266–70. doi: 10.1093/nar/gkx965

73. Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database (Oxford).* (2020), 2020 baaa062. doi: 10.1093/database/baaa062

74. ISO. *ISO 3166 — Country Codes.* ISO. Available online at: https://www.iso.org/iso-3166-country-codes.html (Accessed May 11, 2024).

75. UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Res.* (2022) 51:D523–31. doi: 10.1093/nar/gkac1052

76. *Free Online Spreadsheet Software: Excel | Microsoft 365.* Available online at: https://www.microsoft.com/en-us/microsoft-365/excel (Accessed May 9, 2024).

77. *Google Sheets: Online Spreadsheet Editor | Google Workspace.* Available online at: https://www.facebook.com/GoogleDocs/ (Accessed May 9, 2024).

78. *Prism - GraphPad.* Available online at: https://www.graphpad.com/features (Accessed May 9, 2024).

79. Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, et al. The nf-core framework for community-curated bioinformatics pipelines. *Nat Biotechnol.* (2020) 38:276–8. doi: 10.1038/s41587-020-0439-x

80. *Docker: Accelerated Container Application Development* (2022). Available online at: https://www.docker.com/ (Accessed May 10, 2024).

81. *Introduction to Singularity — Singularity container 3.5 documentation.* Available online at: https://docs.sylabs.io/guides/3.5/user-guide/introduction.html (Accessed May 10, 2024).

82. *Docker Hub Container Image Library | App Containerization.* Available online at: https://hub.docker.com/ (Accessed May 10, 2024).

83. Grüning B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat Methods.* (2018) 15:475–6. doi: 10.1038/s41592-018-0046-7

84. JupyterLab. (2024). Available online at: https://github.com/jupyterlab/jupyterlab (Accessed May 10, 2024).

85. Rstudio. (2024). Available online at: https://github.com/rstudio/rstudio (Accessed May 10, 2024).

86. GitHub. GitHub: Let s build from here. (2024). Available online at: https://github.com/ (Accessed May 10, 2024).

87. Malik-Sheriff RS, Glont M, Nguyen TVN, Tiwari K, Roberts MG, Xavier A, et al. BioModels-15 years of sharing computational models in life science. *Nucleic Acids Res.* (2020) 48:D407–15. doi: 10.1093/nar/gkz1055

88. *Hugging Face – The AI community building the future* (2024). Available online at: https://huggingface.co/ (Accessed May 10, 2024).

89. Benchling. *Cloud-based platform for biotech R&D.* Benchling. (2024). Available online at: https://www.benchling.com/ (Accessed May 11, 2024).

90. L7 Informatics. *Home.* L7 Informatics. Available online at: https://l7informatics.com/ (Accessed May 11, 2024).

91. Zager MA. *Cirro | Data Science Simplified.* Available online at: https://cirro.bio/./ (Accessed May 14, 2024).

92. DNAnexus®. *The Precision Health Data Cloud.* DNAnexus®. Available online at: https://www.dnanexus.com (Accessed May 11, 2024).

93. LatchBio. *LatchBio.* LatchBio. (2024). Available online at: https://latch.bio (Accessed May 11, 2024).

94. *Terra.* Available online at: https://app.terra.bio/ (Accessed May 14, 2024).

95. *Learn more about UK Biobank* (2023). Available online at: https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank (Accessed May 14, 2024).

96. *precisionFDA - Overview.* Available online at: https://precision.fda.gov/ (Accessed May 14, 2024).

97. National Institutes of Health - NIH. *All of Us Research Program | National Institutes of Health (NIH). Research Program.* (2020). Available online at: https://allofus.nih.gov/future-health-begins-all-us (Accessed May 11, 2024).

98. Monica J. *Runaway Cost – Navigating the Cloud.* Thinkwgroup (2024). Available online at: https://www.thinkwgroup.com/runaway-cost/ (Accessed May 15, 2024).

99. Amazon Web Services, Inc. *Cloud Computing Services - Amazon Web Services (AWS).* Amazon Web Services, Inc. (2024). Available online at: https://aws.amazon.com/ (Accessed May 11, 2024).

100. *Cloud Computing Services.* Google Cloud. Available online at: https://cloud.google.com/.

101. Microsoft Azure. *Cloud Computing Services.* (2024). Available online at: https://azure.microsoft.com/en-us (Accessed May 11, 2024).

102. FAQ. *RosettaCommons.* Available online at: https://www.rosettacommons.org/about/faqfaq4 (Accessed May 14, 2024).

103. Ackerman ME, Moldt B, Wyatt RT, Dugast AS, McAndrew E, Tsoukas S, et al. A robust, high-throughput assay to determine the phagocytic activity of clinical antibody samples. *J Immunol Methods.* (2011) 366:8–19. doi: 10.1016/j.jim.2010.12.016

104. Brown EP, Dowell KG, Boesch AW, Normandin E, Mahan AE, Chu T, et al. Multiplexed Fc array for evaluation of antigen-specific antibody effector profiles. *J Immunol Methods.* (2017) 443:33–44. doi: 10.1016/j.jim.2017.01.010

105. Qualtrics. Qualtrics XM - Experience Management Software. (2024). Available online at: https://www.qualtrics.com/ (Accessed May 11, 2024).

106. Workspace G. *Google Forms: Online Form Builder for Business.* Google Workspace. Available online at: https://workspace.google.com/products/forms/.