FEBS Letters

FEBS PRESS science publishing by scientists

THE SCIENTISTS' FORUM

Sharing biological data: why, when, and how

Samantha L. Wilson[1], Gregory P. Way[2], Wout Bittremieux[3,4], Jean-Paul Armache[5], Melissa A. Haendel[6] and Michael M. Hoffman[1,7,8]

1 Princess Margaret Cancer Centre, University Health Network, Toronto, ON, Canada
2 Imaging Platform, Broad Institute of MIT and Harvard, Cambridge, MA, USA
3 Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, USA
4 Department of Computer Science, University of Antwerp, Antwerpen, Belgium
5 Department of Biochemistry & Molecular Biology, The Huck Institutes of Life Sciences, Pennsylvania State University, University Park, PA, USA
6 University of Colorado Anschutz Medical School, Aurora, CO, USA
7 Department of Medical Biophysics, Department of Computer Science, University of Toronto, Toronto, ON, Canada
8 Vector Institute, Toronto, ON, Canada

Data sharing is an essential element of the scientific method, imperative to ensure transparency and reproducibility. Researchers often reuse shared data for meta-analyses or to accompany new data. Different areas of research collect fundamentally different types of data, such as tabular data, sequence data, and image data. These types of data differ greatly in size and require different approaches for sharing. Here, we outline good practices to make your biological data publicly accessible and usable, generally and for several specific kinds of data.

FAIR principles

Sharing data proves more useful when others can easily find and access, interpret, and reuse the data. To maximize the benefit of sharing your data, follow the findable, accessible, interoperable, and reusable (FAIR) guiding principles of data sharing [1] (Box 1), which optimize reuse of generated data. The FAIR principles outline clear standards for ensuring that others can find and access your data and that once accessed, users can easily understand and reuse the data. The FAIR principles provide a clear collection of important details to include within your data and metadata (see 'Data metadata and documentation').

The repositories and practices we recommend below fulfill some of these principles and make it easier for you to follow others. This will not only help others using your data, but can also save you time in the future (see 'The benefits of sharing data to individual researchers').

The National Institutes of Health (NIH), Canadian Institutes of Health Research (CIHR), Monarch Initiative [2,3], and the Research Data Alliance (https://www.rd-alliance.org/) all recommend FAIR principles for data sharing. Amendments to these recommendations that add measures for traceability (such as evidence and provenance), licensing, and connectedness (such as identifiers and versioning) further improve data reusability [4,5].

Why share?

The benefits of sharing data to science and society

Sharing data allows for transparency in scientific studies and allows one to fully understand what occurred in an analysis and reproduce the results. Without complete data, metadata (see 'Data and metadata'), and information about resources used to generate the data, reproducing a study proves impossible [6,7].

Within the biological sciences, we have a problem of data waste—ostensibly shared data that no one ever uses. Many otherwise useful datasets go underused because researchers cannot effectively reuse the data. The inability to reuse arises from lack of discoverability, lack of important information provided, inconsistencies in data and metadata, and licensing issues.

When shared effectively, we can multiply the benefits of large datasets that cost large amounts of funds and research time. Combining previously shared biological data accelerates development of analytical methods used to analyze biological data. Reusing rare samples increases the sample impact. Combining data together in meta-analyses increases study power. Data sharing also leads to fewer duplicate studies. Researchers can build on previous studies to corroborate or falsify their findings rather than repeating the same experiment. Many research projects rely on data from resources such as the Encyclopedia of DNA Elements (ENCODE) Project [8,9]. The

Box 1. FAIR data sharing principles

The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.

F1. (Meta)data are assigned a globally unique and persistent identifier
F2. Data are described with rich metadata (defined by R1 below)
F3. (Meta)data clearly and explicitly include the identifier of the data they describe
F4. (Meta)data are registered or indexed in a searchable resource

Once the user finds the required data, she/he needs to know how can they be accessed, possibly including authentication and authorization.

A1. (Meta)data are retrievable by their identifier using a standardized communications protocol

A1.1 The protocol is open, free, and universally implementable
A1.2 The protocol allows for an authentication and authorization procedure, where necessary

A2. Metadata are accessible, even when the data are no longer available

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
I2. (Meta)data use vocabularies that follow FAIR principles
I3. (Meta)data include qualified references to other (meta)data

The ultimate goal of FAIR is to optimize the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

R1. (Meta)data are richly described with a plurality of accurate and relevant attributes

R1.1 (Meta)data are released with a clear and accessible data usage license
R1.2 (Meta)data are associated with detailed provenance
R1.3 (Meta)data meet domain-relevant community standards

By GO FAIR [1] (https://www.go-fair.org/fair-principles/), provided under the Creative Commons Attribution 4.0 International license.

existence of a large collection of accessible data also aids in the development of cross-cutting analyses such as recount2 [10].

Published manuscripts with reusable data will garner more citations and have more long-term impact on scientific knowledge [11]. As such, many funders now require that grant proposals include a data management and sharing plan describing biological data and metadata [12,13]. Many journals have also implemented policies making public data sharing a requirement upon publication.

The benefits of sharing data to individual researchers

Sharing data increases the impact of a researcher's work and reputation for sound science [14]. Awards for those with an excellent record of data sharing [15] (https://researchsymbionts.org/) or data reuse [16] (https://researchparasite.com/) can exemplify this reputation.

Demonstrating a track record of excellence in resource sharing benefits you when applying for funding. A commitment to and detailed plan for sharing

data publicly increase the perception of a grant proposal's impact [14]. A detailed data sharing plan outlines the types of data you will share, available metadata, and in which repositories you will deposit the data.

Preparing to share data publicly reduces unintentional errors within your own research group. When preparing the data for sharing, providing detailed metadata and documentation will eliminate guesswork, lost details, and maintain tacit knowledge that might otherwise remain unrecorded. Posting data on public repositories with links to the publication and links to data deposited within your publication ensure findability of your data.

Data citation standards now allow directly citing datasets in journal reference list [17]. Citable datasets provide an important incentive to data sharing since those using your shared data can now properly attribute citations to your dataset.

Addressing common concerns about data sharing

Despite the clear benefits of sharing data, some researchers still have concerns about doing so. Some worry that sharing data may decrease the novelty of their work and their chance to publish in prominent journals. You can address this concern by sharing your data only after publication. You can also choose to preprint your manuscript when you decide to share your data. Furthermore, you only need to share the data and metadata required to reproduce your published study.

Time spent on sharing data

Some have concerns about the time it takes to organize and share data publicly. Many add 'data available upon request' to manuscripts instead of depositing the data in a public repository in hopes of getting the work out sooner. It does take time to organize data in preparation for sharing, but sharing data publicly may save you time. Sharing data in a public repository that guarantees archival persistence means that you will not have to worry about storing and backing up the data yourself.

You can consider putting off data sharing tasks as incurring a form of 'sharing debt', by analogy with the concept of technical debt used in software engineering. Delaying these tasks may appear to save you time in the short run, but sharing the data later will take at least as much time as doing it now. You may also incur interest, as it can take longer in the long run to handle individual requests for data availability. Taking a few hours now to organize data and submit it to a repository will save you much of this time.

Human subject data

Sharing of data on human subjects requires special ethical, legal, and privacy considerations. Existing recommendations [18–24] largely aim to balance the privacy of human participants with the benefits of data sharing by de-identifying human participants and obtaining consent for sharing. Sharing human data poses a variety of challenges for analysis, transparency, reproducibility, interoperability, and access [18–24].

Sometimes you cannot publicly post all human data, even after de-identification [25]. We suggest three strategies for making these data maximally accessible. First, deposit raw data files in a controlled-access repository, such as the European Genome-phenome Archive (EGA) [26]67. Controlled-access repositories allow only qualified researchers who apply to access the data. Second, even if you cannot make individual-level raw data available, you can make as much processed data available as possible. This may take the form of summary statistics such as means and standard deviations, rather than individual-level data. Third, you may want to generate simulated data distinct from the original data but statistically similar to it. Simulated data would allow others to reproduce your analysis without disclosing the original data or requiring the security controls needed for controlled access [21].

Data, metadata, and documentation

Data and metadata

Data consist of recorded observations of the biological artifacts or models studied. Metadata describe the primary data and the resources used to generate it.

In a biological context, metadata often provide additional information on samples such as sex, disease, and tissue source site. Metadata often include information about resources such as cell lines and antibodies.

You should share metadata alongside every dataset. A lack of clear metadata for your specific dataset makes it more difficult to understand [27]. This may make it more difficult to reproduce the research or reuse the data. For example, roughly half of >1700 evaluated research studies lacked sufficient specificity in describing resources such as cell lines, organisms, and antibodies to make the study reproducible [6].

In addition to information about samples, metadata also describe experimental protocols and bioinformatic

processes. These include tools used to generate the data, hardware and software versions, processing batch information, and details necessary for understanding data generation.

Most biological disciplines have specific metadata standards that describe the information expected to accompany datasets. For example, genomic researchers have benefited enormously from consistent minimum standard of metadata reporting. The Minimum Information About a Microarray Experiment [28] and Minimum Information About a Next-generation Sequencing Experiment [29] (http://fged.org/projects/minseqe/) guidelines have enabled large-scale efforts to combine and harmonize data, promoting reuse. These guidelines require descriptive standards, experimental design information, essential sample information such as tissue or sex, and bioinformatic processing protocols. Repositories of gene expression data, such as Gene Expression Omnibus (GEO) [26], have mandated use of these guidelines. We discuss metadata standards for individual biological disciplines below.

Using controlled vocabularies or ontologies can improve the rigor of describing biological concepts in your metadata. Ontologies are controlled vocabularies that include both human- and machine-readable semantic relationships between concepts. Widely used biological ontologies include the Gene Ontology [30,31] (http://geneontology.org/) used to annotate gene function and the Uberon anatomy ontology [32] (https://uberon.github.io/). Many repositories or consortium projects require the use of a controlled vocabulary in their metadata standard or data model. For example, the ENCODE Project suggests using Uberon to describe the source of biological tissues.

The formally defined linkages between concepts in an ontology further support interoperability and reusability beyond a simple controlled vocabulary. For example, there exists a logical relationship defining the Gene Ontology term 'dentate gyrus development' (GO:0021542) using a term from Uberon, 'dentate gyrus of hippocampal formation' (UBERON:0001885).

Well-constructed controlled vocabularies and ontologies use globally unique persistent identifiers to refer to each concept. This eliminates ambiguity and makes it easier to link uses of the concept across the whole scientific endeavor. To refer to any controlled vocabulary or ontology term, use a persistent identifier, and version, if applicable.

Documentation

Document your data in three ways: (a) with your manuscript, (b) with description fields in the metadata collected by repositories, and (c) with README files. README files provide abbreviated information about a collection of files. README files associated with biological data should explain organization, file locations, observations and variables present in each file, details on the experimental design, and details on bioinformatic processes.

We regard README files as essential for making your data easy to navigate. Below, we include specific recommendations on README files for different types of biological data.

Source code

Ideally, readers should have all materials needed to completely reproduce the study described in a publication, not just data. These materials include source code, preprocessing, and analysis scripts. Guidelines for organization of computational biology project [33,34] can help you arrange your data and scripts in a way that will make it easier for you and other to access and reuse them.

Licensing

Clear licensing information attached to your data avoids any questions of whether others may reuse it. While copyright law does not protect facts themselves, permission to reuse compilations of facts such as databases may seem less clear without an explicit license. Many data resources turn out not to be as reusable as the providers intended, due to lack of clarity in licensing or restrictive licensing choices [35].

Accompany your data with a license that allows reuse and possibly redistribution. We recommend dedicating your data to the public domain with the CC0 Universal Public Domain Dedication (https://creativecommons.org/choose/zero/). Using CC0 maximizes the ability for others to reuse and remix the data. Other guidelines recommend CC0 [4,36] and many journals and repositories require it.

For nondata artifacts associated with your manuscript, you may wish to use a license with more restrictions than CC0. Relevant licenses include the GNU General Public License (https://www.gnu.org/licenses/gpl-3.0.html) for code and Creative Commons licenses (https://creativecommons.org/choose/) for documents.

When to share

We encourage you to share any data underlying a manuscript by the time of its publication. Many publishers and funding agencies such as NIH [37,38] now

make data sharing an explicit requirement. In addition to sharing all relevant data by publication time, some researchers will go further and make it available when posting a preprint.

Reviewers should have access to underlying data and code when assessing a manuscript [5]. It may seem tempting to restrict data access so that only assigned reviewers can see it during manuscript peer review but this has hidden costs and uncertain benefits. Making the data and code public when submitting the manuscript can avoid this hassle, with few drawbacks. Posting a preprint of the associated manuscript at the same time provides a public record of priority.

How to share: tabular data

Researchers commonly store data in tabular format, an intuitive way to describe multiple similar observations. Tabular format stores information in a structure of rows and columns. Usually, rows contain observations and columns contain variables. In biological data, observations usually refer to samples, replicates, or genes. Variables consist of quantitative or qualitative properties assessed for each observation.

File format

Researchers often save tabular data as spreadsheets. Especially when you have multiple supplementary tables to attach to a manuscript, save the data as a single XLSX workbook [39] with a data dictionary sheet at the beginning of the document. Saving tabular data as XLSX allows for download of all supplementary tables at once. Most programming languages have libraries that make it easy to import and read XLSX workbooks.

Despite the advantages of XLSX workbooks, Microsoft Excel works poorly with certain types of data. Famously, Microsoft Excel changes some gene names to dates [40,41]. This posed a sufficiently severe issue that geneticists changed the gene symbol nomenclature to prevent this mishap [42]. Eluding Excel's mangling of gene symbols can prove complicated. When your data have gene symbols and you have any uncertainty about avoiding corrupting these symbols when saving XLSX workbooks, use non-XLSX formats instead.

When depositing data in public repositories, rather than including it in a manuscript or on the journal's supplementary data Web site, save the data in tab-separated values (TSV) format. This format separates variables with a tab character and separates observations of multiple variables with a newline character. Many programs and programming environments can easily use TSV data.

Avoid comma-separated values (CSV) format, when possible. CSV format has the disadvantage of using commas to separate variables, when commas often occur within variables themselves. This leads to ambiguity and different, incompatible format variants that attempt to solve this problem.

Organization

Certain organizational tactics make data much more interpretable and reduce errors. Broman & Woo [43] and Ellis & Leek [44] provide excellent suggestions on how to organize tabular data. First, ensure that you use the same labels in all areas of your data. For example, inconsistent sex labels, such as 'female', 'Female', 'F', 'f', and '0', make the data hard to read and to reanalyze. Second, pick one representation of data nomenclature and remain consistent throughout your data and documentation. Third, ensure that you use consistent missing value notation, such as 'NA'. Fourth, avoid using spaces in file and column names as this complicates use in many analyses [44]. Incorporating these recommendations makes your data easily interpretable and usable by yourself and others.

Data dictionary

Data dictionaries have a crucial role in organizing your data, especially explaining the variables and their representation. When using XLSX workbooks, add a data dictionary as a separate sheet. When using TSV files, add an additional TSV file containing the data dictionary. Your data dictionaries should provide short names for each variable, a longer text label for the variable, a definition for each variable, data type (such as floating-point number, integer, or string), measurement units, and expected minimum and maximum values. Data dictionaries can make explicit what future users would otherwise have to guess about the representation of data.

Where to share

Share the tabular data most important for interpreting your manuscript as a table within the manuscript itself. You can supply more voluminous data or data less crucial for interpretation as supplementary data attached to the manuscript. Sharing data through the manuscript publisher this way can have three limitations. First, a publisher may limit the size of data you

can include. Second, publishers may make the data difficult to download, especially to download many datasets at once. Third, sometimes publishers have misplaced supplementary data making it difficult to access later or have placed it behind a paywall. To avoid these problems, share especially larger or more complex tabular data in generalist repositories such as Zenodo (https://zenodo.org/; see 'How to share: everything else').

How to share: genomics

File format

Genomic data come in many formats with many different associated biological and technical variables. Usually, raw genomic data consist of sequences stored in FASTA [45] (https://faculty.virginia.edu/wrpearson/fasta/) or FASTQ format [46]. When possible, deposit raw data in CRAM format [47] (https://samtools.github.io/hts-specs/), with unaligned reads included. CRAM files contain sequence information, similar to binary alignment/map (BAM) [48] files, but take up much less space [47] than either a BAM file or a FASTQ file. With unaligned reads included, you should have the ability to reproduce a FASTQ file from a CRAM file.

When possible, deposit your processed data as CRAM, browser extensible data (BED) [49] (https://genome.ucsc.edu/FAQ/FAQformat.html), or TSV files. Format data with genomic regions as BED files instead of generic TSV files. BED files store genomic coordinates of genomic region of interest in the first three columns. The BED format allows additional annotations in subsequent columns, making BED files great for working with genes, binned windows, CpG sites, or transcript data, such as experimental results from genomic assays such as RNA-seq [50–53], chromatin immunoprecipitation-sequencing (ChIP-seq) [54] and assay for transposase-accessible chromatin (ATAC-seq) [55] Using BED formats makes it easy to perform quick analyses on your data with software such as BEDTools [56] (https://bedtools.readthedocs.io/) or Bioawk (https://github.com/lh3/bioawk). Use the bedGraph [57] variant of BED when saving continuous-value data in track format.

In microarray analyses, use CEL (Affymetrix, Santa Clara, CA, USA; https://www.affymetrix.com/support/developer/powertools/changelog/gcos-agcc/cel.html) or IDAT [58] (Illumina, San Diego, CA, USA) file formats for raw data. For storing processed microarray data, store information about genomic regions such as transcripts or CpG sites in BED format.

Compression

Compress large-scale genomic data to minimize the amount of computational storage used. Use gzip (https://www.gnu.org/software/gzip/) compression for single files and ZIP archives for collections of files. Text-based file formats easily compress.

Reference assemblies

Most genomic data have coordinates defined by alignment to a reference genome for a species. Note the reference genome assembly version you align your data to in your manuscript and README file (see 'Documentation'). With advancement of sequencing technologies, genomic coordinates will vary between reference assemblies. For example, a number of parts of the genome changed coordinates between the GRCh37/hg19 [59] and GRCh38/hg38 [60] genome assemblies. Thus, without knowing the reference assembly used for an aligned file, the genomic coordinates hold little value.

Unfortunately, some file formats, such as BED, do not require reference genome assembly metadata. In these cases, make sure to explicitly note which reference assembly you used to align your samples.

Where to share

Public repositories make datasets easily findable by interested parties (Table 1). The GEO [26] repository (https://www.ncbi.nlm.nih.gov/geo/) houses public gene expression and gene regulation data [61]. This includes data on DNA methylation, histone modifications, chromatin organization, and interactions between the genome and proteins such as transcription factors. The submission form requires you to specify both data files and relevant metadata, such as experimental details. After successful deposition, GEO provides you with an accession number for your manuscript. You can place an embargo on your data to withhold public access until publication of your manuscript. GEO will allow an embargo of up to 3 years, but you can change the release date at any time.

Deposit high-throughput sequencing reads that do not fit into GEO in the Sequence Read Archive (SRA) [62] (https://www.ncbi.nlm.nih.gov/sra/). GEO will actually submit raw data files to SRA on your behalf, so you need not submit to both.

Deposit data that contain purely DNA or RNA sequence, rather than quantitative data, in GenBank [63] (https://www.ncbi.nlm.nih.gov/genbank/). These data include sequence of genomic DNA, mRNA, noncoding RNA, plasmids, and synthetic constructs.

Table 1. Genomic repositories

Repository	Purpose	Formats
GEO [26]	Quantitative gene expression, gene regulation, and epigenomic data, including data from RNA-seq [50–53], ChIP-seq [54], Hi-C [118], bisulfite sequencing [119], and microarrays	CRAM [47], BAM [48], SFF, HDF5, FASTQ, bedGraph, bigBed, WIG, bigWig, general feature format (GFF), gene transfer format (GTF), GEOarchive
SRA [62]	Unassembled, high-throughput sequencing reads	CRAM [47], BAM [48], SFF, HDF5, FASTQ
EGA [67]	All kinds of genomics data that contain private genetic or phenotype information on human participants	CRAM [47], BAM [48],, FASTQ, VCF, SFF, HDF5
GenBank [63]	Other DNA and RNA sequences	FASTA

GenBank and the SRA make up part of the International Nucleotide Sequence Database Collaboration (INSDC) [64] (https://www.insdc.org/), which also includes DNA Data Bank of Japan (DDBJ) [65] (https://www.ddbj.nig.ac.jp/) and European Nucleotide Archive [66] (https://www.ebi.ac.uk/ena/). The INSDC members take data submitted to any of these repositories and automatically make it available in the others.

For sensitive genetic and phenotypic information from human participants, EGA [67] (https://ega-archive.org/), a controlled-access genomics archive only permits qualified researchers you approve to access the data. Each dataset must have an associated data access committee that approves access requests and ensures responsible use of the data [25].

How to share: proteomics

File format

Like genomic data, mass spectrometry proteomic experiments generate both raw data and processed data. You should share both. Raw data typically come in a proprietary vendor file format, such as .raw (Thermo Scientific, Waltham, MA, USA), .wiff (SCIEX, Framingham, MA, USA), or .d (Agilent, Santa Clara, CA, USA). Besides the raw data in their original format, also share peak files in the standard mzML file format [68] (https://www.psidev.info/mzML).

Processed data include (a) identification results consisting of peptide-spectrum matches and protein identifications, and (b) quantification results consisting of determined amounts for the identified proteins. Public repositories, such as the ProteomeXchange consortium [69], require raw data and identification data for 'complete' submissions. Provide identification data and quantification data in the standard mzTab format [70] (https://www.psidev.info/mztab). Storing proteomic data in this format, a TSV variant, allows for use of various programming languages without the use of specialized libraries.

Also share other essential files besides the data itself used during the analysis. These include FASTA files with protein sequences or spectral libraries used for spectrum identification.

Metadata and documentation

Provide a README with comprehensive metadata about the experiment, including sample metadata (such as organism and tissues), technical metadata (such as instrument model), and experimental design (such as number of technical and biological replicates). Use the Sample and Data Relationship Format for Proteomics (https://github.com/bigbio/proteomics-metadata-standard) [71] to encode this information in a structured fashion.

Use free-text metadata to describe the study, the sample processing protocol, and the data processing protocol. Comprehensively describe all sample processing steps, including full analytical details. Provide full information on the bioinformatic tools used to process the data, including tool names, version numbers, the organism name, and version information of the FASTA files used for spectrum identification. Also provide the details of any statistical tests and thresholds employed.

For a reanalysis, describe the tools used and how the results differ from the originally deposited data. Do this both in a free-text metadata and in a README document.

Where to share

The ProteomeXchange consortium [69] (https://www.proteomexchange.org/), which includes the main proteomics data repositories such as Proteomics Identifications Database (PRIDE) [72] (https://www.ebi.ac.uk/pride/) and Mass Spectrometry Interactive Virtual

Table 2. Mass spectrometry proteomic repositories

Repository	Purpose
PRIDE [72]	Archival of all kinds of proteomic data
MassIVE	Archival and reanalysis of all kinds of proteomic data
PASSEL [73]	Targeted selected reaction monitoring (SRM) proteomic data
Panorama Public [74]	Targeted proteomic data analyzed using Skyline [120]

Environment (MassIVE) (https://massive.ucsd.edu/), provides a centralized system for sharing mass spectrometry proteomic data (Table 2). To submit data to ProteomeXchange member repositories, you must specify the data type of all files and link raw files to their corresponding peak files and identification results. By default, ProteomeXchange makes submitted datasets private and you can wait until publication time to make the data public. You can include a username and password in scientific manuscripts so that manuscript reviewers can still access the data.

Most ProteomeXchange member repositories take any kind of mass spectrometry proteomics data, whereas some focus on a specific type of data. For example, PeptideAtlas SRMexperiment library (PASSEL) [73] (http://www.peptideatlas.org/passel/) and Panorama Public [74] (https://panoramaweb.org/) only accept deposition of targeted proteomic data.

Some repositories, including MassIVE, store the results of reanalysis of publicly available datasets also. MassIVE makes deposition of data reanalyses simple, as it does not require re-uploading original raw data files already available in public repositories. MassIVE will automatically link the new results to the original data.

How to share: microscopy

Microscopy image data use large amounts of disk space. Microscopy images also have complex associated metadata with great heterogeneity across datasets. The extreme heterogeneity comes from many sources, both biological and technical. Biologists acquire images in two or three spatial dimensions, and sometimes across time via live cell imaging experiments. Biologists also acquire images at different magnifications and across multiple light wavelength channels. The biological substrate captured varies in size (x, y) and depth (z). Biological substrates range from single molecules to whole organisms. Sample preparation before image acquisition also varies widely. Different biologists acquire images using different microscopes with different settings, often using proprietary software and file formats to save output image data with different resolutions, bit depths, and colors. These complexities pose unique data sharing challenges [75,76].

We provide guidelines on how to share microscopy images, intermediate data types, and metadata (Fig. 1). These guidelines have three distinct themes:

1 Use standardized file formats.
2 Select an appropriate repository.
3 Share high-value intermediate data and data processing pipelines.

Following these guidelines will enable the use of your microscopy data in secondary analyses, which will increase the impact of your data.

Compression

Always share images with at least lossless compression. Lossless compression uses less disk space but loses no information as one can expand the compressed file into something identical to the original. Lossy compression, by contrast, loses information.

For very large microscopy datasets, using lossy compression may provide storage and access benefits without losing much vital biological information [77]. Biologists often cringe at losing image resolution or information, but if the loss only marginally decreases analysis performance while increasing access speed and decreasing cost, only sharing the compressed formats may prove the best option. While microscopy data repositories currently offer high ceilings for dataset size (Table 3), this may change as microscopy image datasets grow in size and velocity.

Intermediate data

To maximize the value and impact of your microscopy studies, also share high-value intermediate processed data such as illumination-corrected images. You must correct for uneven illumination around the edges of each microscopy field of view, called shading or vignetting, using computational tools before measuring intensity-related continuous phenotype [78,79]. Typically, additional downstream analyses will use these adjusted images instead of the raw images. Ask the repository if it requires image adjustments before submission.

Image analysis

Depending on experimental goals and strategies, you can also apply an image analysis pipeline. Image analysis produces summary data describing the images,

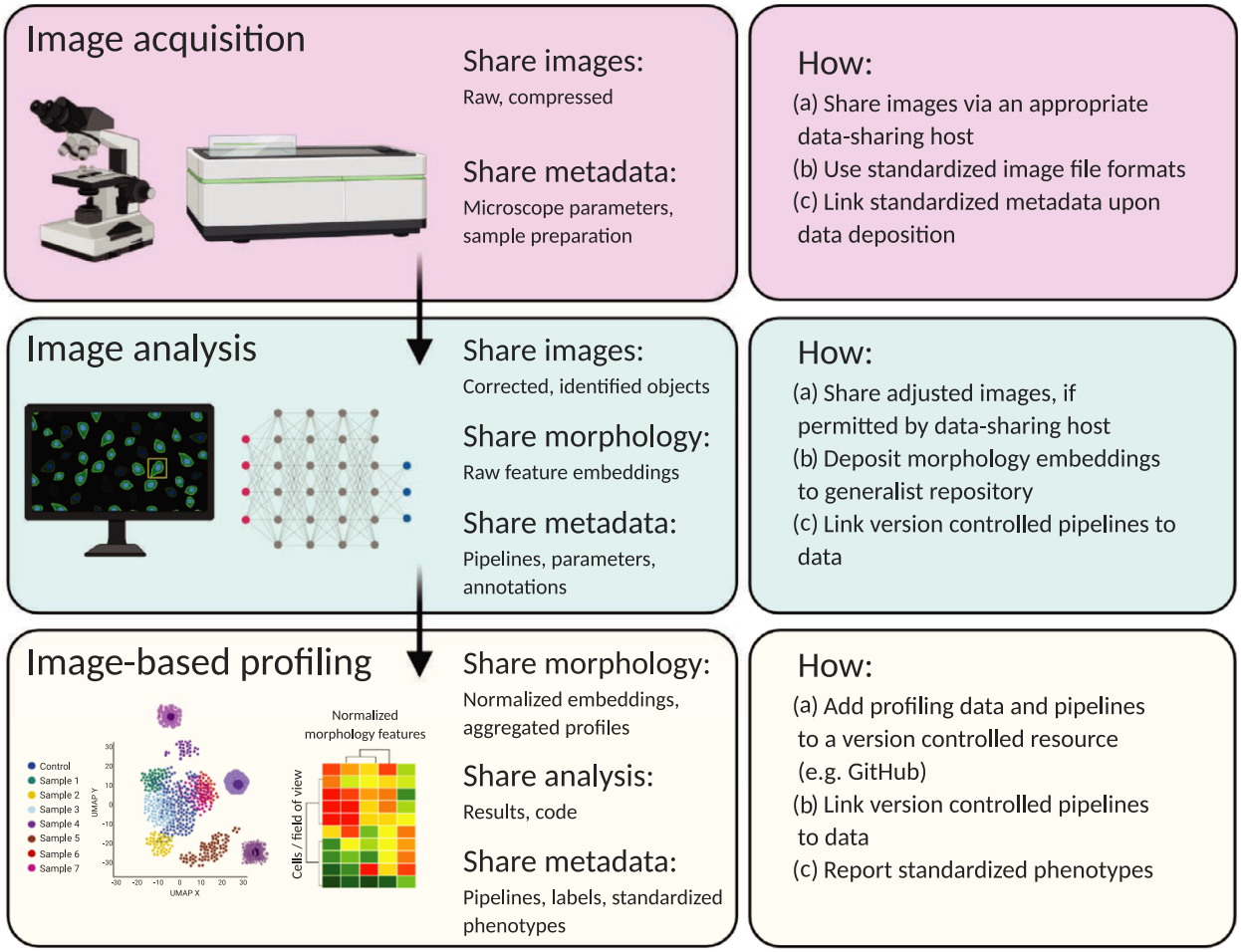

Fig. 1. You should share raw images, corrected images, image-based readouts, analysis pipelines, and comprehensive metadata. While sometimes challenging, this process will ultimately increase the value of your data. Figure made with BioRender (Toronto, ON, Canada).

Table 3. Microscopy image repositories

Repository	Purpose	Substrate	Maximum size	Formats
IDR [100]	Large and complete benchmark microscopy image datasets associated with a publication	Cells and tissues	1000 GB, but you can ask to increase limit	Any Bio-Formats [90], OME-TIFF preferred
EMPIAR [101]	Electron microscopy image data	High-resolution subcellular structures	Tens of TB	TIFF, HDF5, MRC, MRCS, DM4, IMAGIC, SPIDER, FEI
BioImage Archive [102]	Link microscopy image data to associated publications	All nonmedical images not suitable for IDR or EMPIAR	Tens of TB	Any Bio-Formats [90], OME-TIFF preferred
Cell Image Library [103]	Cell images and movies	Cells and intracellular structures	Tens of TB	Any Bio-Formats [90], OME-TIFF preferred
SSBD [104]	Analysis of experimental and computationally simulated biological image data	Any microscopic biological entity, from single molecules to organelles and cells	Tens of TB	Any Bio-Formats [90], OME-TIFF preferred

such as morphology feature embeddings. These summary data take much less disk space than the original images. You can extract this summary data either manually or with specialized software.

Manual annotations provide a gold standard for benchmarking many computational approaches. Researchers generally create such annotations only for small image subsets, and these annotations include few phenotypic measurements [80]. Nevertheless, if you create such annotations, you should make them public. When doing so, include important metadata such as the images used to derive the annotation, the annotator, time collected, and annotation batch (see 'How to share: tabular data').

To more rapidly and consistently measure a richer phenotypic landscape in larger datasets, avoid manual annotation and instead use a computational image analysis pipeline. Many free software packages perform image analysis and extract measurements, including CellProfiler [81], ImageJ [82], Icy [83], PhenoRipper [84], Wndcharm [85], and EBImage [86]. These tools can perform many analyses, including segmenting and counting cells exhibiting a specific phenotype, identifying colocalization of molecules with fluorescent tags, and measuring cell morphology in an unbiased fashion [87].

Image-based profiles

Following image analysis, certain experiments result in high-dimensional readouts that require additional data processing. In these experiments, one extracts image-based profiles. Image-based profiles lack specificity for any target biology—instead, the profiles have no bias toward any biological hypothesis and represent the samples' morphological states. This approach has evolved into a field known as image-based profiling, in which scientists discover biological insights through the aggregation and normalization of morphology features derived from image analysis tools [87–89].

Metadata

To share microscopy data, first catalog experimental metadata in a standard format. Well-structured metadata provide a vital ingredient enabling others to find and use your data [90].

Metadata standardization initiatives provide guidelines on what metadata to share. For example, the Open Microscopy Environment (OME) data model has proposed generic standards and developed software, such as Bio-Formats [90], for standardized metadata reporting and interoperable output file formats [91]. The 4D Nucleome [92] Imaging Standards Working group extended these guidelines to promote rigorous data sharing standards [93].

Individual research communities have augmented these general standards. For example, communities have produced specialized guidelines for reporting cell migration data [94], time-lapse data [95], 3D microscopy images of whole brains (https://www.doryworkspace.org/), and fluorescence microscopy [96].

Follow reporting standards for describing cell phenotypes [97] and cell behavior [98]. To increase the interoperability and value of your data, annotate your images using consistent ontologies.

Where to share

Deposit your data in an appropriate repository [99] (Table 3). Each microscopy data repository has a focused purpose, and accepts data that meet certain size, format, and biological sample conditions. For example, Image Data Resource (IDR) [100] (https://idr.openmicroscopy.org/) accepts benchmark datasets with likely future secondary data analyses and in additional data integration efforts. Electron Microscopy Public Image Archive (EMPIAR) [101] (https://www.ebi.ac.uk/pdbe/emdb/empiar/) accepts high-resolution images from subcellular compartments and biological structures. BioImage Archive [102] (https://www.ebi.ac.uk/bioimage-archive/) provides a home for all other microscopy image datasets, often those of a smaller size. Cell Image Library [103] (http://www.cellimagelibrary.org) hosts a wide variety of biological images and movies for research and education purposes. The Systems Science of Biological Dynamics (SSBD) database [104] (http://ssbd.qbic.riken.jp) also hosts a variety of images and even provides a home for computationally simulated microscopy images.

To determine the appropriate repository, align your microscopy image dataset to the repository with the best-aligned purpose, biological substrate, file size, and output file format (Table 3). When in doubt, contact the repository to determine the suitability of your data. Together, the repositories described here provide a home for all microscopy image datasets.

Depositing your images only in a journal Portable Document Format (PDF) file or pasted in a Microsoft Word or PowerPoint document does not satisfy the FAIR principles (Box 1). This practice will result in low-quality images and compression artifacts and will make future analysis impossible. Do not share data by shipping physical storage devices to requesters or by using cloud provider links [105]. Do not share data

using a custom solution either (see 'How not to share: do not use custom, in-house solutions').

For sharing image data, we usually do not recommend using generalist repositories such as Figshare and Zenodo (see 'How to share: everything else'). These repositories store data that do not have domain-specific resources. They therefore lack the special focus necessary to sufficiently catalog the complexities of microscopy images. Image-based profiles, which consist in small, intermediate data representing morphology feature embeddings, provide the only exception to this. For now, generalist repositories serve as the best place to deposit image-based profiles.

How to share: structural biology

Structural biology encompasses a range of different techniques, including X-ray crystallography, NMR, and multiple kinds of electron microscopy (EM) methods, such as single-particle cryogenic electron microscopy (cryo-EM), cryogenic electron tomography (cryo-ET) [106,107], and microcrystal electron diffraction [108]. Each technique derives information from distinct initial raw data using unique processing approaches.

Where to share

The various structural biology techniques exhibit vast differences in raw data types, files sizes, and paths toward final results. As such, each scientific community developed independent repositories for storing the input and output of these experiments (Table 4). In addition to sharing the final atomic coordinates, each structural biology field developed an individual path to sharing raw and processed data.

Protein Data Bank (PDB) [109] (https://rcsb.org/) serves as a repository for atomic coordinates of nucleic acids, proteins, and larger assemblies. Most journals require structural biology manuscripts to include unique PDB identifiers. Upon deposition of the finalized coordinate files and metadata, authors obtain an unique PDB identifier. Deposition involves creation of a unique identifier and a password. This keeps the files and metadata visible only to the authors and the database operators.

You can place an embargo on your PDB deposition to withhold public access until publication of your manuscript, or 1 year, whichever comes first. We recommend immediate access at the time of publication. Moreover, some journals currently also require coordinate files at the time of submission or by reviewer request. We encourage you to make all your data accessible upon acceptance of the manuscript.

Associate X-ray crystallography structure factors directly with your PDB entries. Store raw diffraction data in the Integrated Resource for Reproducibility in Macromolecular Crystallography (IRRMC) [110,111] (https://www.proteindiffraction.org/).

Deposit NMR structural ensembles in the Biological Magnetic Resonance Bank (BMRB) [112] (https://bmrb.io/). The BMRB usually represents multiple chains under a single identifier. For many experiments, you can deposit raw NMR data, in the form of restraints, in the NMR Restraints Grid [113] (https://restraintsgrid.bmrb.io/).

The most important element of sharing EM data consists in the deposition of the raw, unprocessed data into EMPIAR [101] (https://www.ebi.ac.uk/pdbe/emdb/empiar/). Modern direct detection cameras can generate thousands of images each day. Each image file can contain tens or hundreds of movie frames. Depending on

Table 4. Structural biology repositories

Repository	Purpose	Substrate	Maximum size	Formats
PDB [109]	Atomic coordinates and ensembles	Subcellular structures	Tens of MB	PDB, mmCIF
EMDB [117]	3D reconstructions from processed EM data	Subcellular structure		MRC, CCP4
EMPIAR [101]	EM raw and processed image data	Subcellular structures (single-particle cryo-EM, cryo-ET)	Tens of TB	TIFF, HDF5, MRC, MRCS, DM4, EER, IMAGIC, SPIDER, SCIPION, EMDB-SFF, AMIRA, STL, VTK, VTP, OBJ, AVI, JPEG, PNG, EMX, BLENDER, TXT
IRRMC [110,111]	X-ray diffraction raw data	Subcellular structures	Hundreds of MB	Raw diffraction data formats
BMRB [112]	NMR raw data	Subcellular structures	GB	CCPN, mmCIF, PDB, NMR-STAR, X-PLOR

the file format, raw data from a single session can range from 1 TB to 10 TB. Deposit the raw, uncorrected movie stacks, as well as summed, motion-corrected micrographs, final particle coordinates, and final alignment files in EMPIAR. This greatly simplifies data validation and reproducibility. It also simplifies software development—having access to raw and processed training data improves heterogeneity classification and machine learning approaches [114–116].

Cryo-EM microscopy uses Coulomb potential densities to build and refine coordinate files. Thus, the final coordinate model depends on the quality of the EM reconstructions, on your individual choices, and on the model-refinement approaches used. Thus, providing final filtered and unfiltered maps has great importance to the validation claims made in a manuscript. Deposit the coordinates and final calculated Coulomb potential maps in the Electron Microscopy Data Bank (EMDB) [117] (https://www.ebi.ac.uk/pdbe/emdb/), obtaining unique PDB and EMDB identifiers.

When a cryo-EM dataset reveals structural variability, provide a consensus output and deposit all the associated models and maps needed to support the claims of the manuscript in separate depositions. In addition, if multiple 3D variability clusters result in a large number of intermediate maps, add them to an EMPIAR deposition, along with the raw data. As new technologies enable collecting more data in shorter times, the focus on describing motion will increase. Accordingly, computational structural heterogeneity analysis approaches will become more sophisticated.

How to share: everything else

For some types of data not covered above, no specialized repositories exist. Deposit these kinds of data in generalist repositories that can manage many different types of data.

Organization

Organize your data depositions with raw data separate from results. Use ZIP archives to collect your data so that viewers can preview individual files in repositories such as Zenodo. To make your data clear and interpretable, include a README with a detailed description of the project and an explanation of what each of the files contains.

Where to share

First, see whether an appropriate data repository exists in the re3data directory (https://www.re3data.org/). For cases where no such repository exists, we recommend Zenodo (https://zenodo.org/), a generalist repository that allows for deposition of data, code, analysis, and manuscripts and has robust semantic versioning as well as a persistence guarantee.

Open Science Framework (OSF) (https://osf.io/) provides a system for organizing scientific projects, including data, code, and protocols. It also serves as a generalist repository, allowing you to share data and other materials simply by making you OSF project publicly available.

How not to share: Do not use custom, in-house solutions

Hosting your data using a customized solution you create may seem attractive. For example, some share their data using public Amazon Web Services Simple Storage Service (S3) links or even building a new repository specifically for their project. Using a custom solution provides an illusion of complete control. In reality, custom efforts usually result in something fragile with uncertain permanence and in difficulty for tracking attribution and citation.

Do not reinvent the wheel. Third-party repositories have more permanence, exist outside the control of the original data generators, and provide storage and infrastructure maintenance cost savings. Third-party repositories also enforce metadata standards that facilitate FAIR sharing principles (Box 1). These repositories have access to funding streams and institutional commitments that individual investigators lack. Users interact with third-party repositories frequently, and many have had negative experiences by custom hosting efforts, which generate more problems than solutions.

Discussion

We suggest a four-step checklist for biological researchers to complete when submitting a manuscript:

1 Deposit raw and processed data. Use a specialist repository if possible. Dedicate these datasets to the public domain with CC0.
2 Deposit code to a generalist repository.
3 Deposit all miscellaneous files to generalist repository.
4 Put all repository accession numbers and license information in your manuscript.

We encourage you to create a laboratory publication checklist that contains the necessary steps for a laboratory member to prepare a manuscript and associated

artifacts for publication. Use the four-step checklist as a starting point, and add details specific to you and the kinds of data you work with.

We understand that some will find the above recommendations difficult or overwhelming at first. We encourage you to do what you can. Improving your data management and sharing practices gradually will still provide great value for you and other researchers. Finally, the intent to share matters.

Acknowledgements

We thank Erin Weisbart (0000-0002-6437-2458) and Anne E. Carpenter (0000-0003-1555-8261) (Broad Institute) for helpful discussions on how to share microscopy and image data. This work was supported by the Natural Sciences and Engineering Research Council of Canada (RGPIN-2015-03948 to MMH), the CIHR Fellowship (MFE-171256 to SLW), the Research Foundation—Flanders (12W0418N to WB), and the NIH (R24OD011883 to MAH).

Author contributions

SLW and MMH conceptualized the publication; SLW and MMH involved in project administration; GPW designed the visualization for the publication; SLW, GPW, WB, and J-PA wrote—original draft; SLW, GPW, MAH, and MMH wrote—review and editing.

References

1 Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018.

2 McMurry JA, Köhler S, Washington NL, Balhoff JP, Borromeo C, Brush M, Carbon S, Conlin T, Dunn N, Engelstad M *et al.* (2016) Navigating the phenotype frontier: the Monarch Initiative. *Genetics* **203**, 1491–1495.

3 Shefchek KA, Harris NL, Gargano M, Matentzoglu N, Unni D, Brush M, Keith D, Conlin T, Vasilevsky N, Zhang XA *et al.* (2020) The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res* **48**, D704–D715.

4 Haendel M, Su A, McMurry J, Chute CG, Mungall C, Good B, Wu C, McWeeney S, Hochheiser H, Robinson P *et al.* FAIR-TLC: Metrics to Assess Value of Biomedical Digital Repositories: Response to RFI NOT-OD-16-133. https://doi.org/10.5281/zenodo.203295

5 McMurry JA, Juty N, Blomberg N, Burdett T, Conlin T, Conte N, Courtot M, Deck J, Dumontier M, Fellows DK *et al.* (2017) Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. *PLOS Biol* **15**, e2001414.

6 Vasilevsky NA, Brush MH, Paddock H, Ponting L, Tripathy SJ, LaRocca GM and Haendel MA (2013) On the reproducibility of science: unique identification of research resources in the biomedical literature. *PeerJ* **1**, e148.

7 Ford J (2013) Unreliable research: trouble at the lab. *Economist* **409**, 26–31.

8 ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74.

9 ENCODE Project Consortium, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, Adrian J, Kawli T, Davis CA, Dobin A *et al.* (2020) Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710.

10 Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, Hansen KD, Jaffe AE, Langmead B and Leek JT (2017) Reproducible RNA-seq analysis using recount2. *Nat Biotechnol* **35**, 319–321.

11 Piwowar HA, Day RS and Fridsma DB (2007) Sharing detailed research data is associated with increased citation rate. *PLOS One* **2**, e308.

12 National Institutes of Health (2020) Final NIH Policy for Data Management and Sharing. NIH Guide to Grants and Contracts, NOT-OD-21–013. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html

13 National Institutes of Health (2020) Supplemental information to the NIH Policy for Data Management and Sharing: selecting a repository for data resulting from NIH-supported research. NIH Guide to Grants and Contracts, NOT-OD-21–016. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-016.html

14 Pierce HH, Dev A, Statham E and Bierer BE (2019) Credit data generators for data reuse. *Nature* **570**, 30–32.

15 Byrd JB and Greene CS (2017) Data-sharing models. *N Engl J Med* **376**, 2305.

16 Greene CS, Garmire LX, Gilbert JA, Ritchie MD and Hunter LE (2017) Celebrating parasites. *Nat Genet* **49**, 483–484.

17 Starr J, Castro E, Crosas M, Dumontier M, Downs RR, Duerr R, Haak LL, Haendel M, Herman I, Hodson S *et al.* (2015) Achieving human and machine accessibility of cited data in scholarly publications. *PeerJ Comput Sci* **1**, e1.

18 Knoppers BM and Thorogood AM (2017) Ethics and big data in health. *Curr Opin Syst Biol* **4**, 53–57.

19 Council for International Organizations of Medical Sciences (2016). International Ethical Guidelines for

Health-related Research Involving Humans. 4th edn. Council for International Organizations of Medical Sciences (CIOMS), Geneva, Switzerland. https://cioms.ch/wp-content/uploads/2017/01/WEB-CIOMS-EthicalGuidelines.pdf
20 Clayton EW, Evans BJ, Hazel JW and Rothstein MA (2019) The law of genetic privacy: applications, implications, and limitations. *J Law Biosci* **6**, 1–36.
21 Haibe-Kains B, Adam GA, Hosny A, Khodakarami F, Waldron L, Wang B, McIntosh C, Goldenberg A, Kundaje A, Greene CS *et al.* (2020) Transparency and reproducibility in artificial intelligence. *Nature* **586**, E14–E16.
22 Zook M, Barocas S, Boyd D, Crawford K, Keller E, Gangadharan SP, Goodman A, Hollander R, Koenig BA, Metcalf J *et al.* (2017) Ten simple rules for responsible big data research. *PLOS Comput Biol* **3**, e1005399.
23 Deverka PA, Majumder MA, Villanueva AG, Anderson M, Bakker AC, Bardill J, Boerwinkle E, Bubela T, Evans BJ, Garrison NA *et al.* (2017) Creating a data resource: what will it take to build a medical information commons? *Genome Med* **9**, 84.
24 Malin B, Goodman K *et al.* (2018) Between access and privacy: challenges in sharing health data. *Yearb Med Inform* **27**, 55–59.
25 Byrd JB, Greene AC, Prasad DV, Jiang X and Greene CS (2020) Responsible, practical genomic data sharing that accelerates research. *Nat Rev Genet* **21**, 615–629.
26 Clough E and Barrett T (2016) The Gene Expression Omnibus database. *Methods Mol Biol* **1418**, 93–110.
27 Leipzig J, Nüst D, Hoyt CT, Soiland-Reyes S, Ram K and Greenberg J (2020) The role of metadata in reproducible computational research. *arXiv* [PREPRINT].
28 Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC *et al.* (2001) Minimum Information About a Microarray Experiment (MIAME)—toward standards for microarray data. *Nat Genet* **29**, 365–371.
29 Brazma A (2009) Minimum Information About a Microarray Experiment (MIAME)—successes, failures, challenges. *Sci World* **9**, 420–423.
30 Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat Genet* **25**, 25–29.
31 Gene Ontology Consortium (2021) The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res* **49**, D325–D334.
32 Haendel MA, Balhoff JP, Bastian FB, Blackburn DC, Blake JA, Bradford Y, Comte A, Dahdul WM, Dececchi TA, Druzinsky RE *et al.* (2014) Unification of multi-species vertebrate anatomy ontologies for comparative biology in Uberon. *J Biomed Semantics* **5**, 21.
33 Noble WS (2009) A quick guide to organizing computational biology projects. *PLOS Comput Biol* **5**, e1000424.
34 Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L and Teal TK (2017) Good enough practices in scientific computing. *PLOS Comput Biol* **13**, e1005510.
35 Carbon S, Champieux R, McMurry JA, Winfree L, Wyatt LR and Haendel MA (2019) An analysis and metric of reusable data licensing practices for biomedical resources. *PLOS One* **14**, e0213090.
36 Schofield PN, Bubela T, Weaver T, Portilla L, Brown SD, Hancock JM, Einhorn D, Tocchini-Valentini G, de Angelis MH and Rosenthal N (2009) Post-publication sharing of data and tools. *Nature* **461**, 171–173.
37 National Institute of Allergy and Infectious Diseases (2019) Data sharing for grants—final research data SOP. In Research Rules & Policies. https://www.niaid.nih.gov/research/grants-data-sharing-final-research
38 National Institute of Allergy and Infectious Diseases (2017) Genomic data sharing plan examples. In Genomic Data Sharing. https://www.niaid.nih.gov/research/gds-plan-examples
39 International Organization for Standardization, International Electrotechnical Commission (2016) ISO/IEC 29500-1:2016: Information Technology — Document Description and Processing Languages — Office Open XML File Formats — Part 1: Fundamentals and Markup Language Reference. International Organization for Standardization, Geneva.
40 Zeeberg BR, Riss J, Kane DW, Bussey KJ, Uchio E, Linehan WM, Barrett JC and Weinstein JN (2004) Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics. *BMC Bioinformatics* **5**, 80.
41 Ziemann M, Eren Y and El-Osta A (2016) Gene name errors are widespread in the scientific literature. *Genome Biol* **17**, 177.
42 Bruford EA, Braschi B, Denny P, Jones TE, Seal RL and Tweedie S (2020) Guidelines for human gene nomenclature. *Nat Genet* **52**, 754–758.
43 Broman KW and Woo KH (2018) Data organization in spreadsheets. *Am Stat* **72**, 2–10.
44 Ellis SE and Leek JT (2018) How to share data for collaboration. *Am Stat* **72**, 53–57.
45 Lipman DJ and Pearson WR (1985) Rapid and sensitive protein similarity searches. *Science* **227**, 1435–1441.
46 Cock PJ, Fields CJ, Goto N, Heuer ML and Rice PM (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res* **38**, 1767–1771.

47 Fritz MH-Y, Leinonen R, Cochrane G and Birney E (2011) Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res* **21**, 734–740.
48 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G and Durbin R (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079.
49 Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM and Haussler D (2002) The human genome browser at UCSC. *Genome Res* **12**, 996–1006.
50 Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH and Ecker JR (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133**, 523–536.
51 Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J and Bähler J (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453**, 1239–1243.
52 Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G *et al.* (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods* **5**, 613–619.
53 Mortazavi A, Williams BA, McCue K, Schaeffer L and Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621–628.
54 Johnson DS, Mortazavi A, Myers RM and Wold B (2007) Genome-wide mapping of *in vivo* protein-DNA interactions. *Science* **316**, 1497–1502.
55 Buenrostro JD, Giresi PG, Zaba LC, Chang HY and Greenleaf WJ (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**, 1213–1218.
56 Quinlan AR and Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842.
57 Karolchik D, Hinrichs AS and Kent WJ (2012) The UCSC genome browser. *Curr Protoc Bioinformatics* **40**, 1.4.1–1.4.33.
58 Smith ML, Baggerly KA, Bengtsson H, Ritchie ME and Hansen K (2013) illuminaio: an open source IDAT parsing tool for Illumina microarrays. *F1000Research* **2**, 264.
59 Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, Chen H-C, Agarwala R, McLaren WM, Ritchie GR *et al.* (2011) Modernizing reference genome assemblies. *PLOS Biol* **9**, e1001091.
60 Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen H-C, Kitts PA, Murphy TD, Pruitt KD, Thibaud-Nissen F, Albracht D *et al.* (2017) Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res* **27**, 849–864.
61 Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M *et al.* (2012) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res* **41**, D991–D995.
62 Kodama Y, Shumway M and Leinonen R (2012) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res* **40**, D54–D56.
63 Sayers EW, Cavanaugh M, Clark K, Pruitt KD, Schoch CL, Sherry ST and Karsch-Mizrachi I (2020) GenBank. *Nucleic Acids Res* **49**, D92–D96.
64 Arita M, Karsch-Mizrachi I, Cochrane G, International Nucleotide Sequence Database Collaboration (2021) The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res* **49**, D121–D124.
65 Fukuda A, Kodama Y, Mashima J, Fujisawa T and Ogasawara O (2021) DDBJ update: streamlining submission and access of human data. *Nucleic Acids Res* **49**, D71–D75.
66 Harrison PW, Ahamed A, Aslam R, Alako BT, Burgin J, Buso N, Courtot M, Fan J, Gupta D, Haseeb M *et al.* (2021) The European Nucleotide Archive in 2020. *Nucleic Acids Res* **49**, D82–D85.
67 Lappalainen I, Almeida-King J, Kumanduri V, Senf A, Spalding JD, Saunders G, Kandasamy J, Caccamo M, Leinonen R, Vaughan B *et al.* (2015) The European Genome-phenome Archive of human data consented for biomedical research. *Nat Genet* **47**, 692–695.
68 Martens L, Chambers M, Sturm M, Kessner D, Levander F, Shofstahl J, Tang WH, Römpp A, Neumann S, Pizarro AD *et al.* (2011) mzML—a community standard for mass spectrometry data. *Mol Cell Proteomics* **10**, R110.000133.
69 Deutsch EW, Bandeira N, Sharma V, Perez-Riverol Y, Carver JJ, Kundu DJ, García-Seisdedos D, Jarnuczak AF, Hewapathirana S, Pullman BS *et al.* (2019) The ProteomeXchange consortium in 2020: enabling 'big data' approaches in proteomics. *Nucleic Acids Res* **48**, D1145–D1152.
70 Griss J, Jones AR, Sachsenberg T, Walzer M, Gatto L, Hartler J, Thallinger GG, Salek RM, Steinbeck C, Neuhauser N *et al.* (2014) The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. *Mol Cell Proteomics* **13**, 2765–2775.
71 Perez-Riverol Y, European Bioinformatics Community for Mass Spectrometry (2020) Toward a sample metadata standard in public proteomics repositories. *J Proteome Res* **19**, 3906–3909.
72 Perez-Riverol Y, Csordas A, Bai J, Bernal-Llinares M, Hewapathirana S, Kundu DJ, 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.
73 Farrah T, Deutsch EW, Kreisberg R, Sun Z, Campbell DS, Mendoza L, Kusebauch U, Brusniak M-Y, Hüttenhain R, Schiess R *et al.* (2012) PASSEL: the PeptideAtlas SRMexperiment Library. *Proteomics* **12**, 1170–1175.
74 Sharma V, Eckels J, Schilling B, Ludwig C, Jaffe JD, MacCoss MJ and MacLean B (2018) Panorama Public: a public repository for quantitative data sets processed in Skyline. *Mol Cell Proteomics* **17**, 1239–1244.
75 Zaritsky A (2018) Sharing and reusing cell image data. *Mol Biol Cell* **29**, 1274–1280.
76 Marqués G, Pengo T and Sanders MA (2020) Imaging methods are vastly underreported in biomedical research. *eLife* **9**, e55133.
77 Balázs B, Deschamps J, Albert M, Ries J and Hufnagel L (2017) A real-time compression library for microscopy images. *bioRxiv* 164624 [PREPRINT].
78 Singh S, Bray M-A, Jones T and Carpenter A (2014) Pipeline for illumination correction of images for high-throughput microscopy. *J Microsc* **256**, 231–236.
79 Peng T, Thorn K, Schroeder T, Wang L, Theis FJ, Marr C and Navab N (2017) A BaSiC tool for background and shading correction of optical microscopy images. *Nat Commun* **8**, 14836.
80 Ljosa V, Sokolnicki KL and Carpenter AE (2012) Annotated high-throughput microscopy image sets for validation. *Nat Methods* **9**, 637.
81 McQuin C, Goodman A, Chernyshev V, Kamentsky L, Cimini BA, Karhohs KW, Doan M, Ding L, Rafelski SM, Thirstrup D *et al.* (2018) CellProfiler 3.0: next-generation image processing for biology. *PLOS Biol* **16**, e2005970.
82 Rueden CT, Schindelin J, Hiner MC, DeZonia BE, Walter AE, Arena ET and Eliceiri KW (2017) ImageJ2: ImageJ for the next generation of scientific image data. *BMC Bioinformatics* **18**, 529.
83 De Chaumont F, Dallongeville S, Chenouard N, Hervé N, Pop S, Provoost T, Meas-Yedid V, Pankajakshan P, Lecomte T, Montagner YL *et al.* (2012) Icy: an open bioimage informatics platform for extended reproducible research. *Nat Methods* **9**, 690–696.
84 Rajaram S, Pavie B, Wu LF and Altschuler SJ (2012) PhenoRipper: software for rapidly profiling microscopy images. *Nat Methods* **9**, 635–637.
85 Shamir L, Orlov N, Eckley DM, Macura T, Johnston J and Goldberg IG (2008) Wndchrm—an open source utility for biological image analysis. *Source Code Biol Med* **3**, 13.
86 Pau G, Fuchs F, Sklyar O, Boutros M and Huber W (2010) EBImage—an R package for image processing with applications to cellular phenotypes. *Bioinformatics* **26**, 979–981.
87 Caicedo JC, Cooper S, Heigwer F, Warchal S, Qiu P, Molnar C, Vasilevich AS, Barry JD, Bansal HS, Kraus O *et al.* (2017) Data-analysis strategies for image-based cell profiling. *Nat Methods* **14**, 849–863.
88 Scheeder C, Heigwer F and Boutros M (2018) Machine learning and image-based profiling in drug discovery. *Curr Opin Syst Biol* **10**, 43–52.
89 Chandrasekaran SN, Ceulemans H, Boyd JD and Carpenter AE (2020) Image-based profiling for drug discovery: due for a machine-learning upgrade? *Nat Rev Drug Discov* **20**, 1–15.
90 Linkert M, Rueden CT, Allan C, Burel J-M, Moore W, Patterson A, Loranger B, Moore J, Neves C, MacDonald D *et al.* (2010) Metadata matters: access to image data in the real world. *J Cell Biol* **189**, 777–782.
91 Goldberg IG, Allan C, Burel J-M, Creager D, Falconi A, Hochheiser H, Johnston J, Mellen J, Sorger PK and Swedlow JR (2005) The Open Microscopy Environment (OME) Data Model and XML file: open tools for informatics and quantitative analysis in biological imaging. *Genome Biol* **6**, R47.
92 Dekker J, Belmont AS, Guttman M, Leshyk VO, Lis JT, Lomvardas S, Mirny LA, Oshea CC, Park PJ, Ren B *et al.* (2017) The 4D Nucleome project. *Nature* **549**, 219–226.
93 Huisman M, Hammer M, Rigano A, Farzam F, Gopinathan R, Smith C, Grunwald D and Strambio-De-Castillia C (2020) Minimum Information guidelines for fluorescence microscopy: increasing the value, quality, and fidelity of image data. *arXiv* [PREPRINT].
94 Gonzalez-Beltran AN, Masuzzo P, Ampe C, Bakker G-J, Besson S, Eibl RH, Friedl P, Gunzer M, Kittisopikul M, Le Dévédec SE *et al.* (2020) Community standards for open cell migration data. *GigaScience* **9**, giaa041.
95 Burek P, Scherf N and Herre H (2019) Ontology patterns for the representation of quality changes of cells in time. *J Biomed Semantics* **10**, 16.
96 Lee J-Y and Kitaoka M (2018) A beginner's guide to rigor and reproducibility in fluorescence imaging experiments. *Mol Biol Cell* **29**, 1519–1525.
97 Jupp S, Malone J, Burdett T, Heriche J-K, Williams E, Ellenberg J, Parkinson H and Rustici G (2016) The Cellular Microscopy Phenotype Ontology. *J Biomed Semantics* **7**, 28.
98 Sluka JP, Shirinifard A, Swat M, Cosmanescu A, Heiland RW and Glazier JA (2014) The Cell Behavior Ontology: describing the intrinsic biological behaviors of real and model cells seen as active agents. *Bioinformatics* **30**, 2367–2374.
99 Dance A (2020) Find a home for every imaging data set. *Nature* **579**, 162–163.

100 Williams E, Moore J, Li SW, Rustici G, Tarkowska A, Chessel A, Leo S, Antal B, Ferguson RK, Sarkans U *et al.* (2017) The Image Data Resource: a bioimage data integration and publication platform. *Nat Methods* **14**, 775–781.
101 Iudin A, Korir PK, Salavert-Torres J, Kleywegt GJ and Patwardhan A (2016) EMPIAR: a public archive for raw electron microscopy image data. *Nat Methods* **13**, 387–388.
102 Ellenberg J, Swedlow JR, Barlow M, Cook CE, Sarkans U, Patwardhan A, Brazma A and Birney E (2018) A call for public archives for biological image data. *Nat Methods* **15**, 849–854.
103 Orloff DN, Iwasa JH, Martone ME, Ellisman MH and Kane CM (2013) The cell: an image library-CCDB: a curated repository of microscopy data. *Nucleic Acids Res* **41**, D1241–D1250.
104 Tohsato Y, Ho KHL, Kyoda K and Onami S (2016) SSBD: a database of quantitative data of spatiotemporal dynamics of biological phenomena. *Bioinformatics* **32**, 3471–3479.
105 Andreev A and Koo DE (2020) Practical guide to storage of large amounts of microscopy data. *Microsc Today* **28**, 42–45.
106 Grimm R, Bärmann M, Häckl W, Typke D, Sackmann E and Baumeister W (1997) Energy filtered electron tomography of ice-embedded actin and vesicles. *Biophys J* **72**, 482–489.
107 Dierksen K, Typke D, Hegerl R, Walz J, Sackmann E and Baumeister W (1995) Three-dimensional structure of lipid vesicles embedded in vitreous ice and investigated by automated electron tomography. *Biophys J* **68**, 1416–1422.
108 Shi D, Nannenga BL, Iadanza MG and Gonen T (2013) Three-dimensional electron crystallography of protein microcrystals. *eLife* **2**, e01345.
109 Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN and Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242.
110 Grabowski M, Langner KM, Cymborowski M, Porebski PJ, Sroka P, Zheng H, Cooper DR, Zimmerman MD, Elsliger MA, Burley SK and *et al.* (2016) A public database of macromolecular diffraction experiments. *Acta Crystallogr D Struct Biol* **72**, 1181–1193.
111 Grabowski M, Cymborowski M, Porebski PJ, Osinski T, Shabalin IG, Cooper DR and Minor W (2019) The Integrated Resource for Reproducibility in Macromolecular Crystallography: experiences of the first four years. *Struct Dyn* **6**, 064301.
112 Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z *et al.* (2008) L. BioMagResBank. *Nucleic Acids Res* **36**, 402–408.
113 Doreleijers JF, Nederveen AJ, Vranken W, Lin J, Bonvin AMJJ, Kaptein R, Markley JL and Ulrich EL (2005) BioMagResBank databases DOCR and FRED containing converted and filtered sets of experimental NMR restraints and coordinates from over 500 protein PDB structures. *J Biomol NMR* **32**, 1–12.
114 Wagner T, Merino F, Stabrin M, Moriya T, Antoni C, Apelbaum A, Hagel P, Sitsel O, Raisch T, Prumbaum D *et al.* (2019) SPHIRE-crYOLO is a fast and accurate fully automated particle picker for cryo-EM. *Commun Biol* **2**, 218.
115 Bepler T, Kelley K, Noble AJ and Berger B (2020) Topaz-Denoise: general deep denoising models for cryoEM and cryoET. *Nat Commun* **11**, 5208.
116 Tegunov D and Cramer P (2019) Real-time cryo-electron microscopy data preprocessing with Warp. *Nat Methods* **16**, 1146–1152.
117 Lawson CL, Baker ML, Best C, Bi C, Dougherty M, Feng P, van Ginkel G, Devkota B, Lagerstedt I, Ludtke SJ *et al.* (2011) EMDataBank.org: unified data resource for CryoEM. *Nucleic Acids Res* **39**, D456–D464.
118 Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293.
119 Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, Molloy PL and Paul CL (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci USA* **89**, 1827–1831.
120 MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, Frewen B, Kern R, Tabb DL, Liebler DC and MacCoss MJ (2010) Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966–968.

Correspondence
M. M. Hoffman, Princess Margaret Cancer Centre, University Health Network, Princess Margaret Cancer Research Tower 11-311, 101 College St, Toronto, ON M6J 2X3, Canada
Tel: +1 415 5871 7481
E-mail: michael.hoffman@utoronto.ca