Sharing images

Images are among the richest data types that biologists collect, yet most biological images are not available for reanalysis or reuse. This may be changing.

It is now standard practice to make available the data underlying the conclusions of a biological study. Gene expression and genomic variation data, the coordinates that underpin protein or chemical structural models, mass spectrometry-based proteomics data—all these data types are deposited in dedicated repositories at publication. This is typically required by journals and expected by researchers. These practices in turn enable reanalysis of the data by other scientists.

Not so for imaging data. In particular for fluorescence microscopic data on cells, tissues, and developing organisms, data deposition in a repository is the exception rather than the rule.

This is reflected in journal policies; in contrast to requirements for genomic or structural data, most journals, including the Nature Research journals, do not mandate deposition of image data. The *Journal of Cell Biology* DataViewer makes it possible for authors to permit the download of raw images from published papers, but JCB policy does not require image data deposition.

There are certainly exceptions to the dearth of image data sharing. Specific imaging projects—some large-scale screens or cellular or tissue atlases, for example—often do make image data available. At *Nature Methods*, authors reporting a new computational method implemented as software are asked to provide sample data on which their software can be tested, and these data are typically provided as part of the online supplement. So-called generalist repositories like figshare, Dryad, and BioStudies also can and do host image data sets and have the capacity for at least medium-sized data sets—tens to even hundreds of gigabytes—with reportedly little trouble. But unlike for genomic data or, for that matter, for medical and magnetic resonance images, there is no dedicated large-scale biological imaging data repository.

The multidimensionality, variety, lack of standardized metadata, and size of imaging data undoubtedly all contribute to this state of affairs. Perhaps for the same reasons there has simply not been a culture of data sharing in imaging, certainly not at the level prevalent in genomics.

One may also ask whether the scientific rationale for image sharing is as clear as it is for other data types. Of course, verifying and building upon the conclusions in a study is as important for imaging projects as for any other project. Making cellular images in a study more comprehensively available should improve reproducibility, since this would reduce the frequency with which conclusions may be drawn based on cherry-picked data, exemplified by an image of a single cell or well. But a strong rationale for data sharing also typically rests on whether the data can be used to ask questions not posed by the original study. That image-based atlases could be productively reused is clear. But for small-scale imaging studies asking a particular research question, the potential for reuse is less obvious.

Arguing for the scientific value of image data availability, a paper in this issue (p775) presents the Image Data Resource (IDR), which hosts a selection of annotated image data sets with standardized metadata and with integrated visualization and analysis tools. Not strictly a repository, IDR is rather a platform for cross-dataset analysis and, as such, illustrates the value of making such data available. Although IDR is populated with just 35 data sets (most of them from image-based screens), phenotypes are reported for 90% of known genes in at least three experimental contexts. The authors show that analysis of phenotypes across data sets suggests network representations that could not be proposed from individual studies.

Image data sets also serve as benchmarks for comparing performance of analytical methods. When they include prior annotation (identified cells or cellular structures, segmented nuclei, tracked lineages), such data sets are invaluable for training algorithms in a variety of analytical tasks. As resources for the development and comparison of methods, the value of large amounts of available image data cannot be overstated.

Nevertheless, given the likely costs of storing research images comprehensively, the prospect of a dedicated image repository raises more questions than answers. Should all images underlying biological experiments be stored, or only a subset? Which subset? Will present data formats suffice, or are new formats or a consensus format needed? Are the numerical data underlying images more important than the images themselves? Alternatively, should the emphasis be on storing processed data or analytical annotations? The answers depend on the uses to which we imagine such data will be put.

As data integration becomes more common and also more powerful, it would be unfortunate if the rich information in image data was not available for such efforts. We invite our readers to consider making their image data available and to communicate with us their views regarding the potential value or inconveniences of this practice.