



University of Dundee

The Image Data Resource

Williams, Eleanor; Moore, Josh; Li, Simon W.; Rustici, Gabriella; Tarkowska, Aleksandra; Chessel, Anatole

Published in:
Nature Methods

DOI:
[10.1038/nmeth.4326](https://doi.org/10.1038/nmeth.4326)

Publication date:
2017

Document Version
Peer reviewed version

[Link to publication in Discovery Research Portal](#)

Citation for published version (APA):

Williams, E., Moore, J., Li, S. W., Rustici, G., Tarkowska, A., Chessel, A., Leo, S., Antal, B., Ferguson, R. K., Sarkans, U., Brazma, A., Carzo Salas, R. E., & Swedlow, J. R. (2017). The Image Data Resource: a bioimage data integration and publication platform. *Nature Methods*, 14(8), 775-781. <https://doi.org/10.1038/nmeth.4326>

General rights

Copyright and moral rights for the publications made accessible in Discovery Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

1 **The Image Data Resource: A Bioimage Data Integration and Publication Platform**

2

3

4

5 Eleanor Williams^{1,2*}, Josh Moore^{1*}, Simon W. Li^{1*}, Gabriella Rustici¹, Aleksandra
6 Tarkowska¹, Anatole Chessel^{3,6}, Simone Leo^{1,5}, Bálint Antal³, Richard K. Ferguson¹, Ugis
7 Sarkans², Alvis Brazma², Rafael E. Carazo Salas^{3,4}, Jason R. Swedlow^{1,7}

8

9

10 ¹Centre for Gene Regulation & Expression & Division of Computational Biology, University of
11 Dundee, Dundee, Scotland, UK;

12

13 ²European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, United
14 Kingdom

15

16 ³Pharmacology & Genetics Departments and Cambridge Systems Biology Centre, University
17 of Cambridge, Cambridge, UK

18

19 ⁴School of Cell and Molecular Medicine, University of Bristol, Bristol, UK

20

21 ⁵Center for Advanced Studies, Research, and Development in Sardinia (CRS4), Pula(CA),
22 Italy

23

24 ⁶LOB, Ecole Polytechnique, CNRS, INSERM, Université Paris-Saclay, Palaiseau, France

25

26 *equal contribution

27

28 ⁷Correspondence to:

29

30 email: jrswedlow@dundee.ac.uk

31 phone +44 1382 385819

32 fax +44 1382 388072

33 URL: www.openmicroscopy.org

34

35 Abstract

36 **Access to primary research data is vital for the advancement of science. To extend**
37 **the data types supported by community repositories, we built a prototype Image Data**
38 **Resource (IDR) that collects and integrates imaging data acquired across many**
39 **different imaging modalities. IDR links high-content screening, super-resolution**
40 **microscopy, time-lapse and digital pathology imaging experiments to public genetic**
41 **or chemical databases, and to cell and tissue phenotypes expressed using controlled**
42 **ontologies. Using this integration, IDR facilitates the analysis of gene networks and**
43 **reveals functional interactions that are inaccessible to individual studies. To enable**
44 **re-analysis, we also established a computational resource based on Jupyter**
45 **notebooks that allows remote access to the entire IDR. IDR is also an open source**
46 **platform that others can use to publish their own image data. Thus IDR provides both**
47 **a novel on-line resource and a software infrastructure that promotes and extends**
48 **publication and re-analysis of scientific image data.**

49

50 Much of the published research in the life sciences is based on image datasets that sample
51 3D space, time, and the spectral characteristics of detected signal (e.g., photons, electrons,
52 proton relaxation, etc.) to provide quantitative measures of cell, tissue and organismal
53 processes and structures. The sheer size of biological image datasets makes data
54 submission, handling and publication extremely challenging. An image-based genome-wide
55 “high-content” screen (HCS) may contain over a million images, and new “virtual slide” and
56 “light sheet” tissue imaging technologies generate individual images that contain gigapixels
57 of data showing tissues or whole organisms at subcellular resolutions. At the same time,
58 published versions of image data often are mere illustrations: they are presented in
59 processed, compressed formats that cannot convey the measurements and multiple
60 dimensions contained in the original image data and that can no longer be easily subjected
61 to re-analysis. Furthermore, conventional publications neither include the metadata that
62 define imaging protocols, biological systems and perturbations nor the processing and
63 analytic outputs that convert the image data into quantitative measurements.

64

65 Several public image databases have appeared over the last few years. These provide on-
66 line access to image data, enable browsing and visualization, and in some cases include
67 experimental metadata. The Allen Brain Atlas, the Human Protein Atlas, and the Edinburgh
68 Mouse Atlas all synthesize measurements of gene expression, protein localization and/or
69 other analytic metadata with coordinate systems that place biomolecular localization and
70 concentration into a spatial and biological context¹⁻³. Similarly, many other examples of
71 dedicated databases for specific imaging projects exist, each tailored to its aims and its
72 target community⁴⁻⁸. There are also a number of public resources that serve as true
73 scientific, structured repositories for image data, i.e., that collect, store and provide
74 persistent identifiers for long-term access to submitted datasets, as well as provide rich
75 functionalities for browsing, search and query. One archetype is the EMDDataBank, the
76 definitive community repository for molecular reconstructions recorded by electron
77 microscopy⁹. The *Journal of Cell Biology* has built the JCB DataViewer ([http://jcb-](http://jcb-dataviewer.rupress.org/)
78 [dataviewer.rupress.org/](http://jcb-dataviewer.rupress.org/)), which publishes image datasets associated with its on-line
79 publications. The CELL Image Library publishes several thousand community-submitted
80 images, some of which are linked to publications¹⁰. FigShare stores 2D pictures derived from
81 image datasets, and can provide links for download of image datasets (<http://figshare.com>).

82 The EMDataBank recently has released a prototype repository for 3D tomograms, the
83 EMPIAR resource¹¹. Finally, the BioStudies and Dryad archives include support for browsing
84 and downloading image data files linked to studies or publications¹² (<https://datadryad.org/>).
85 Some of these provide a resource for a specific imaging domain (e.g., EMDataBank) or
86 experiment (Mitocheck), while others archive datasets and provide links to a related
87 publication available at an external journal's website (BioStudies). However, no existing
88 resource links several independent biological imaging datasets to provide an “added value”
89 platform, like the Expression Atlas achieves for a broad set of gene expression datasets¹³
90 and UniProt delivers for protein sequence and function datasets¹⁴.

91
92 Inspired by these “added value” resources, we have built a next-generation Image Data
93 Resource (IDR) – an added value platform that combines data from multiple independent
94 imaging experiments and from many different imaging modalities, integrates them into a
95 single resource, and makes the data available for re-analysis in a convenient, scalable form.
96 IDR provides, for the first time, a prototyped resource that supports browsing, search,
97 visualization and computational processing within and across datasets acquired from a wide
98 variety of imaging domains. For each study, metadata related to the experimental design
99 and execution, the acquisition of the image data, and downstream interpretation and
100 analysis are stored in IDR alongside the image data and made available for search and
101 query through a web interface and a single API. Wherever possible, we have mapped the
102 phenotypes determined by dataset authors to a common ontology. For several studies, we
103 have calculated comprehensive sets of image features that can be used by others for re-
104 analysis and the development of phenotypic classifiers. By harmonizing the data from
105 multiple imaging studies into a single system, IDR users can query across studies and
106 identify phenotypic links between different experiments and perturbations.

107 Results

108 Current IDR

109 IDR is currently populated with 24 imaging studies comprising 35 screens or imaging
110 experiments from the biological imaging community, most of which are related to and linked
111 to published works (Table 1). IDR holds ~42 TB of image data in ~36M image planes and
112 ~1M individual experiments, and includes all associated experimental (e.g., genes, RNAi,
113 chemistry, geographic location), analytic (e.g., submitter-calculated image regions and
114 features), and functional annotations. Datasets in human cells
115 (<https://idr.openmicroscopy.org/webclient/?show=well-45407>), *Drosophila*
116 (<https://idr.openmicroscopy.org/webclient/?show=well-547609>) and fungi
117 (<https://idr.openmicroscopy.org/webclient/?show=well-590686>;
118 <https://idr.openmicroscopy.org/webclient/?show=well-469267>), super resolution 3DSIM
119 images of centrosomes (<https://idr.openmicroscopy.org/webclient/?show=dataset-51>);
120 dSTORM images of nuclear pores ([https://idr.openmicroscopy.org/webclient/?show=dataset-
121 61](https://idr.openmicroscopy.org/webclient/?show=dataset-61)), a comprehensive chemical screen in human cells
122 (<https://idr.openmicroscopy.org/webclient/?show=plate-4101>), a live cell screen in human
123 cells (Mitocheck; <https://idr.openmicroscopy.org/webclient/?show=well-771034>) and
124 histopathology whole slide images of tissues from several mouse mutants
125 (<https://idr.openmicroscopy.org/webclient/?show=dataset-369>) are included. Finally, imaging
126 from Tara Oceans, a global survey of plankton and other marine organisms, is also included

127 (<https://idr.openmicroscopy.org/webclient/?show=plate-4751>). The current collection of
128 datasets samples a variety of biomedically-relevant biological processes like cell shape,
129 division and adhesion, at scales ranging from nanometer-scale localization of proteins in
130 cells to millimeter-scale structure of tissues from animals.

131 **Genetic, Chemical and Functional Annotation in IDR**

132 To enable querying across the different datasets stored in IDR, we have included
133 annotations describing experimental perturbations (genetic mutants, siRNA targets and
134 reagents, expressed proteins, cell lines, drugs, etc.) and phenotypes declared by the study
135 authors either from quantitative analysis or visual inspection of the image data. Wherever
136 possible, experimental metadata in IDR link to external resources that are the authoritative
137 resource for those metadata (Ensembl, NCBI, PubChem, etc.).
138

139 The result is that IDR is a sampling of phenotypes related to experimental perturbations
140 across several independent studies. Many of the studies in IDR perturb gene function by
141 mutation or siRNA depletion. To calculate the sampling of gene orthologues, we used
142 Ensembl's BioMart resource¹⁵ to access a normalized list of gene orthologues. Overall,
143 19,601 gene orthologues are sampled, and 84.1% of gene orthologues are sampled more
144 than 20 times. 90.3% of gene orthologues are sampled in three or more studies, so even in
145 this early incarnation the phenotypes of perturbations in the majority of known genes are
146 sampled in several different assays and organisms.

147
148 We have normalized the phenotypes included in studies submitted to IDR. Functional
149 annotations (e.g., "increased peripheral actin") have been converted to defined terms in the
150 Cellular Microscopy Phenotype Ontology (CMPO)¹⁶, or other ontologies in collaboration with
151 the data submitters (e.g., <https://idr.openmicroscopy.org/webclient/?show=image-109846>).
152 Overall, 88% of the functional annotations have links to defined, published controlled
153 vocabularies. 158 different ontology-normalized phenotypes (e.g., "increased number of
154 actin filaments", "mitosis arrested") are included in IDR, and 136 are reported by authors in
155 only one study. Nonetheless, these phenotypes are well-sampled-- the mean number of
156 samples per phenotype, across HCS and other imaging datasets is 698 and the median is
157 144. This skewing occurs because some phenotypes are very common or are over-
158 represented in specific assays, e.g. "protein localized in cytosol phenotype",
159 (CMPO_0000393;
160 https://idr.openmicroscopy.org/mapr/phenotype/?value=CMPO_0000393). Nonetheless,
161 there are several cases where phenotypes are observed in multiple orthogonal assays. Two
162 examples are the "round cell" phenotype (CMPO_0000118;
163 https://idr.openmicroscopy.org/mapr/phenotype/?value=CMPO_0000118) and the
164 "increased nuclear size" phenotype (CMPO_0000140;
165 https://idr.openmicroscopy.org/mapr/phenotype/?value=CMPO_0000140). Figure 1
166 summarizes the sampling of phenotypes across the current IDR datasets. Several classes of
167 phenotypes are included, and many cases are sampled in thousands of experiments. In
168 total, IDR includes >1M individual experiments (Table 1), with ~9 % annotated with
169 experimentally observed phenotypes.

170 **Data Visualization in IDR**

171 IDR integrates image data and metadata from several studies into a single resource. The
172 current IDR web user interface (WUI) is based on the open source OMERO.web

173 application¹⁷ supplemented with a plugin allowing datasets to be viewed by ‘Study’, ‘Genes’,
 174 ‘Phenotypes’, ‘siRNAs’, ‘Antibodies’, ‘Compounds’, and ‘Organisms’ (see Supplementary
 175 Note). Using this architecture makes the integrated data resource available for access and
 176 re-use in several ways. Image data are viewable as thumbnails for each study (e.g.,
 177 <https://idr.openmicroscopy.org/webclient/?show=plate-4349>) and multi-dimensional images
 178 can be viewed and browsed (e.g., <https://idr.openmicroscopy.org/webclient/?show=well-45501>
 179 and <https://idr.openmicroscopy.org/webclient/?show=well-93714>). Tiled whole slide
 180 images used in histopathology are also supported (e.g.,
 181 <https://idr.openmicroscopy.org/webclient/?show=image-1920135>). Where identified regions
 182 of interest (ROIs) have been submitted with the image data, these have been included and
 183 linked, and where possible, made available through the IDR WUI (e.g.,
 184 <https://idr.openmicroscopy.org/webclient/?show=well-590686> and
 185 https://idr.openmicroscopy.org/webclient/img_detail/1230005/). IDR images, thumbnails and
 186 metadata are accessible through the IDR WUI and web-based API in JSON format (see
 187 Supplementary Note). They also can be embedded into other pages using the OMERO.web
 188 gateway (e.g., <https://www.eurobioimaging-interim.eu/image-data-repository.html>).

189 **Standardized Interfaces for Imaging Metadata**

190 IDR integrates imaging data from many different, independent studies. These data were
 191 acquired using several different imaging modalities, in the absence of any over-arching
 192 standards for experimental, imaging or analytic metadata. While efforts like MIACA
 193 (<http://miaca.sourceforge.net/>), NeuroVault¹⁸, MULTIMOT¹⁹ and several other projects have
 194 proposed data standards in specific imaging subdomains, there is not yet a metadata
 195 standard that crosses all of the imaging domains potentially served by IDR. We therefore
 196 sought to adopt lightweight methods from other communities that have had broad
 197 acceptance²⁰ and converted metadata submitted in custom formats – spreadsheets, PDFs,
 198 MySQL databases, and Microsoft Word documents -- into a consistent tabular format
 199 inspired by the MAGE-TAB and ISA-TAB specifications^{21, 22} that could then be used for
 200 importing semi-structured metadata like gene and ontology identifiers into OMERO²³. We
 201 also used the Bio-Formats software library to identify and convert well-defined, semantically-
 202 typed elements that describe the imaging metadata (e.g., image pixel size) as specified in
 203 the OME Data Model^{24, 25}. The resulting translation scripts were used to integrate datasets
 204 from multiple distinct studies and imaging modalities into a single resource. The scripts are
 205 publicly available (see Online Methods) and thus comprise a framework for recognizing and
 206 reading a range of metadata types across several imaging domains into a common, open
 207 specification.

208 **Added Value of IDR**

209 Because IDR links gene names and phenotypes, query results that combine genes and
 210 phenotypes across multiple studies are possible through simple text-based search.
 211 Searching for the gene *SGOL1* (<https://idr.openmicroscopy.org/mapr/gene/?value=SGOL1>)
 212 returns a range of phenotypes from four separate studies associated with mitotic defects (for
 213 example, CMPO_0000118, CMPO_0000305, CMPO_0000212, CMPO_0000344, etc.)^{4, 26}
 214 but also an accelerated secretion phenotype (CMPO_0000246) in a screen for defects in
 215 protein secretion²⁷. A second example is provided in a histopathology study of tissue
 216 phenotypes in a series of mouse mutants. Knockout of carbonic anhydrase 4 (*Car4*;
 217 <https://idr.openmicroscopy.org/mapr/gene/?value=Car4>) in mouse results in a range of
 218 defects in homeostasis in the brain, rib growth and male fertility²⁸⁻³⁰. Data held in IDR show

219 abnormal nuclear phenotypes in several tissues from *Car4*^{-/-} mice (e.g., GI:
220 <https://idr.openmicroscopy.org/webclient/?show=dataset-153>; liver:
221 <https://idr.openmicroscopy.org/webclient/?show=image-1918940>; male reproductive tract:
222 <https://idr.openmicroscopy.org/webclient/?show=image-1918953>). The human orthologue,
223 *CA4*, is involved in certain forms of retinitis pigmentosa^{31, 32}. Data presented in IDR from the
224 Mitocheck study show that siRNA depletion of *CA4* in HeLa cells⁴ also results in abnormally
225 shaped nuclei (<https://idr.openmicroscopy.org/webclient/?show=well-828419>) consistent with
226 a defect in some aspect of the cell division cycle.

227

228 Phenotypes across distinct studies can also be used to build novel representations of gene
229 networks. Figure 2A shows the gene network created when the gene knockouts or
230 knockdowns that caused an elongated cell phenotype (CMPO_0000077) in studies in *S.*
231 *pombe* and human cells are linked by queries to String DB³³ and visualized in Cytoscape³⁴
232 (see Supplementary Note and Supplementary Table 1). The genes discovered in the three
233 studies form interconnected, non-overlapping, complementary networks that connect
234 specific macromolecular complexes to the elongated cell phenotype. For example, *HELZ2*,
235 *MED30*, *MED18* and *MED20* are all part of the Mediator Complex, but were identified as
236 “elongated cell” hits in separate studies using different biological models (idr0001-A,
237 idr0008-B, idr0012-A, Figure 2B). *POLR2G* (from idr0012-A), *PAF1* (from idr0001-A) and
238 *SUPT16H* (from idr0008-B) were scored as elongated cell hits in these studies and are all
239 part of the Elongation complex in the RNA Polymerase II transcription pathway. Finally,
240 *ASH2L* (“elongated cell phenotype” in idr0012-A), associates with *SETD1A* and *SETD1B*
241 (“elongated cell phenotype” in idr0001-A) to form the Set1 histone methyltransferase (HMT).
242 These examples demonstrate that these individual hits are probably not due to off-target
243 effects or characteristics of individual biological models but arise through conserved, specific
244 functions of large macromolecular complexes. This shows the utility and importance of
245 combining phenotypic data of studies from different organisms and scales, and of integrating
246 metadata from independent studies, to generate added value that can enhance the
247 understanding of biological mechanisms and lead to new mechanistic hypothesis and
248 predictions.

249

250 The integration of experimental, image and analytic metadata also provides an opportunity
251 to include new functionalities for more advanced visualization and analytics of imaging data
252 and metadata, bringing further added value to the original studies and datasets. We have
253 added the data analytics tool Mineotaur³⁵ to one of IDR’s datasets
254 (<https://idr.openmicroscopy.org/mineotaur/>). This allows visual querying and analysis of
255 quantitative feature data. For instance, having shown that components of the Set1 HMT
256 function in controlling cell morphology in *S. pombe* and human cells, we noticed that genes
257 like *ASH2L* were in the “elongated cell” network based on human cell data (idr0012-A) but
258 not *S. pombe* data, where *ash2*, the *S. pombe* *ASH2L* orthologue, was not annotated as a
259 cell elongation “hit”. We first noted that *ash2* has a microtubule cytoskeleton phenotype
260 (<https://idr.openmicroscopy.org/webclient/?show=well-592371>). We then queried the criteria
261 previously used for cell shape hits in the Sysgro screen (idr0001-A) and found that *ash2* fell
262 just below the cutoff originally used in this study to define phenotypic hits for cell shape
263 (Supplementary Note). When combined with results on *ASH2L* from HeLa cells (Figure 2B)
264 these results suggest that the Set1 HMT has a strongly conserved role in controlling cell
265 shape and the cytoskeleton in unicellular and multicellular organisms.

266 **Data Integration and Access**

267 Like most modern on-line resources IDR makes data available through a web user-interface
268 as well as a web-based JSON API. This encourages third-parties to make use of IDR in their
269 own sites. For example, image data in IDR has been linked to study data in BioStudies,
270 thereby extending the linkage of study and image metadata (e.g.,
271 <https://www.ebi.ac.uk/biostudies/studies/S-EPMC4704494>), and to PhenolImageShare³⁶, an
272 on-line phenotypic repository (e.g.,
273 [http://www.phenoimageshare.org/search/?term=&hostName=Image+Data+Repository+\(IDR\)](http://www.phenoimageshare.org/search/?term=&hostName=Image+Data+Repository+(IDR))
274). These are examples of use of IDR as a service that delivers data for other applications to
275 integrate and reuse.

276
277 To add further value and extend the possibilities for reuse of IDR data, we are calculating
278 comprehensive sets of feature vectors of IDR image data using the open source tool WND-
279 CHARM³⁷. To date full WND-CHARM features have been calculated for images in idr0002-
280 A, idr0005-A, idr0008-B, idr0009-A, idr0009-B, idr0012-A, and parts of idr0013-A and
281 idr0013-B. Feature calculations for other IDR datasets are in progress. Features are stored
282 in IDR using OMERO's HDF5-based data store and available through the OMERO API (see
283 Supplementary Note).

284
285 The integration of image-based phenotypes and calculated features makes IDR an attractive
286 candidate for computational re-analysis. To ease the access to IDR's TB-scale datasets, we
287 have connected IDR to a Jupyter notebook-based computational resource
288 (<https://idr.openmicroscopy.org/jupyter>) that exposes IDR datasets via an API
289 (<https://idr.openmicroscopy.org/about/api.html>). We include exemplar notebooks that provide
290 visualization of image features using PCA, access to images annotated with CMPO
291 phenotypes, calculation of gene networks, calculation of WND-CHARM features for
292 individual images and recreation of Figures 1 and 2 from IDR data. Alternatively, users can
293 run their own analyses using notebooks stored in GitHub ([https://github.com/IDR/idr-](https://github.com/IDR/idr-notebooks)
294 [notebooks](https://github.com/IDR/idr-notebooks)). To allow re-use of IDR metadata locally, we have made all IDR databases,
295 metadata and thumbnails available for download and have built Ansible scripts that
296 automate the deployment of the IDR software stack (original image data are not included;
297 see Supplementary Note).

298 **Discussion**

299 Making data public and available is a critical part of the scientific enterprise³⁸
300 (<https://wellcome.ac.uk/what-we-do/our-work/expert-advisory-group-data-access>)
301 (<https://royalsociety.org/topics-policy/projects/science-public-enterprise/report/>). To take the
302 next step in facilitating the reuse and meta-analysis of image datasets we have built IDR, a
303 next-generation data technology that integrates and publishes image data and metadata
304 from a wide range of imaging modalities and scales in a consistent format. IDR integrates
305 experimental, imaging, phenotypic and analytic metadata from several independent studies
306 into a single resource, allowing new modes of biological Big Data querying and analysis. As
307 more datasets are added to IDR, they will potentiate and catalyze the generation of new
308 biological hypotheses and discoveries.

309

310 In IDR, we have linked image metadata from several independent studies. Experimental,
311 imaging phenotypic and analytic metadata are recorded in a consistent format. Rather than

312 declaring and attempting to enforce a strict imaging data standard, IDR provides tools for
313 supporting community formats and releases these as a framework that facilitates data reuse.
314 We hope that the availability of this framework will provide incentives for others to structure
315 their metadata in shareable formats that can be read into IDR or other applications, whether
316 based on OMERO or not. In the future, we can imagine that these and other capabilities
317 could be extended in IDR - or similar repositories that link to IDR - to enable systematic
318 integration, visualization and analytics across imaging studies, thereby helping to harness
319 and capitalize on the exponentially increasing amounts of bio-imaging data that the
320 community generates.

321

322 As of this writing, IDR has published 35 reference image datasets grouped into 24 studies
323 (Table 1) and, utilizing EMBL-EBI's Embassy Cloud, has capacity to receive and publish
324 many more. Authors of scientific publications that are already published or under submission
325 can submit accompanying image datasets for publication in IDR, using the metadata
326 specifications and formats we have built. Once published, the datasets can be browsed and
327 viewed through IDR's WUI, or queried and re-analyzed using the IDR computational
328 resource. Details about the submission process are available
329 (<https://idr.openmicroscopy.org/about/submission.html>).

330

331 IDR software and technology is open source, so it can be accessed and built into other
332 image data publication systems. This supports the building of technology and installations
333 that integrate and publish bio-image data for the scientific community, allowing discoveries
334 and predictions similar to what we have shown in Figure 2. IDR therefore functions both as a
335 resource for image data publication and as a technology platform that supports the creation
336 of on-line scientific image databases and services. In the future, those databases and
337 services may ultimately amalgamate to form resources analogous to the genomic resources
338 that are the foundation of much of modern biology.

339

340 Acknowledgements

341 We thank all the study authors who submitted datasets for inclusion in the IDR for their
342 contributions and help in incorporating their datasets. We also thank the systems support
343 team at EMBL-EBI, in particular Richard Boyce, David Ocana, Charles Short, and Andy
344 Cafferkey for their support of the project's use of the Embassy Cloud. We are particularly
345 grateful to Simon Jupp for help with adding and defining new ontology terms. The IDR
346 project was funded by the BBSRC (BB/M018423/1) and Horizon 2020 Framework
347 Programme of the European Union under grant agreement No. 688945 (Euro-Biolmaging
348 Prep Phase II). Updates to OMERO and Bio-Formats were supported by the Wellcome Trust
349 (095931/Z/11/Z) and Horizon 2020 Framework Programme of the European Union under
350 grant agreement No. 634107 (MULTIMOT). RECS was funded by a BBSRC Responsive
351 Mode grant (BB/K006320/1), a European Research Council Starting Researcher Investigator
352 Grant (SYSGRO) and the University of Bristol.

353

354 Author Contributions

355 J.R.S., A.B. and R.E.C.S conceived and funded the project which was overseen by
356 J.R.S. J.M., S.W.L, A.T., S.L. and E.W. built the IDR software stack: J.M. designed the
357 software architecture and managed the software development team; S.W.L. built all the tools
358 for deploying IDR in the OpenStack cloud and ran all the IDR systems; A.T. built Mapr, the
359 metadata querying application; S.L. updated Bio-Formats to read the incoming datasets;
360 E.W. performed all data curation and annotation. G.R., S.W.L. and E.W. sourced and
361 received the datasets. A.C. analyzed features of the integrated datasets. B.A. helped with
362 the IDR/Mineotaur integration. R.K.F. designed the updates to the OMERO UI. U.S. helped
363 with the integration of IDR datasets into BioStudies.

364 Competing Financial Interests

365 J. R. S is affiliated with Glencoe Software, Inc., an open-source US-based commercial
366 company that provides commercial licenses for OME software.

367
368
369

References

370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422

1. Uhlen, M. et al. Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
2. Hawrylycz, M.J. et al. An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* **489**, 391-399 (2012).
3. Armit, C. et al. eMouseAtlas, EMAGE, and the spatial dimension of the transcriptome. *Mammalian genome : official journal of the International Mammalian Genome Society* **23**, 514-524 (2012).
4. Neumann, B. et al. Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. *Nature* **464**, 721-727 (2010).
5. Graml, V. et al. A genomic Multiprocess survey of machineries that control and link cell shape, microtubule organization, and cell-cycle progression. *Dev Cell* **31**, 227-239 (2014).
6. Koh, J.L. et al. CYCLOPs: A Comprehensive Database Constructed from Automated Analysis of Protein Abundance and Subcellular Localization Patterns in *Saccharomyces cerevisiae*. *G3 (Bethesda)* **5**, 1223-1232 (2015).
7. Gonczy, P. et al. Functional genomic analysis of cell division in *C. elegans* using RNAi of genes on chromosome III. *Nature* **408**, 331-336. (2000).
8. Fowlkes, C.C. et al. A quantitative spatiotemporal atlas of gene expression in the *Drosophila* blastoderm. *Cell* **133**, 364-374 (2008).
9. Lawson, C.L. et al. EMDatabank.org: unified data resource for CryoEM. *Nucleic Acids Res.* **39**, D456-464 (2011).
10. Orloff, D.N., Iwasa, J.H., Martone, M.E., Ellisman, M.H. & Kane, C.M. The cell: an image library-CCDB: a curated repository of microscopy data. *Nucleic Acids Res* **41**, D1241-1250 (2013).
11. Iudin, A., Korir, P.K., Salavert-Torres, J., Kleywegt, G.J. & Patwardhan, A. EMPIAR: a public archive for raw electron microscopy image data. *Nat Methods* **13**, 387-388 (2016).
12. McEntyre, J., Sarkans, U. & Brazma, A. The BioStudies database. *Molecular systems biology* **11**, 847 (2015).
13. Petryszak, R. et al. Expression Atlas update--an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Res* **44**, D746-752 (2016).
14. UniProt, C. UniProt: a hub for protein information. *Nucleic Acids Res* **43**, D204-212 (2015).
15. Yates, A. et al. Ensembl 2016. *Nucleic Acids Res* **44**, D710-716 (2016).
16. Jupp, S. et al. The cellular microscopy phenotype ontology. *Journal of biomedical semantics* **7**, 28 (2016).
17. Allan, C. et al. OMERO: flexible, model-driven data management for experimental biology. *Nat Methods* **9**, 245-253 (2012).
18. Gorgolewski, K.J. et al. NeuroVault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. *Frontiers in neuroinformatics* **9**, 8 (2015).
19. Masuzzo, P. et al. An open data ecosystem for cell migration research. *Trends Cell Biol* **25**, 55-58 (2015).
20. Brazma, A., Krestyaninova, M. & Sarkans, U. Standards for systems biology. *Nat Rev Genet* **7**, 593-605 (2006).
21. Sansone, S.A. et al. Toward interoperable bioscience data. *Nat Genet* **44**, 121-126 (2012).
22. Rayner, T.F. et al. A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. *BMC Bioinformatics* **7**, 489 (2006).

- 423 23. Li, S. et al. Metadata management for high content screening in OMERO. *Methods*
424 (2015).
- 425 24. Linkert, M. et al. Metadata matters: access to image data in the real world. *J. Cell.*
426 *Biol.* **189**, 777-782 (2010).
- 427 25. Goldberg, I.G. et al. The Open Microscopy Environment (OME) Data Model and XML
428 File: Open Tools for Informatics and Quantitative Analysis in Biological Imaging.
429 *Genome Biol.* **6**, R47 (2005).
- 430 26. Heriche, J.K. et al. Integration of biological data by kernels on graph nodes allows
431 prediction of new genes involved in mitotic chromosome condensation. *Mol Biol Cell*
432 **25**, 2522-2536 (2014).
- 433 27. Simpson, J.C. et al. Genome-wide RNAi screening identifies human proteins with a
434 regulatory function in the early secretory pathway. *Nat Cell Biol* **14**, 764-774 (2012).
- 435 28. Shah, G.N. et al. Carbonic anhydrase IV and XIV knockout mice: roles of the
436 respective carbonic anhydrases in buffering the extracellular space in brain. *Proc*
437 *Natl Acad Sci U S A* **102**, 16771-16776 (2005).
- 438 29. Scheibe, R.J. et al. Carbonic anhydrases IV and IX: subcellular localization and
439 functional role in mouse skeletal muscle. *Am J Physiol Cell Physiol* **294**, C402-412
440 (2008).
- 441 30. Wandernoth, P.M. et al. Role of carbonic anhydrase IV in the bicarbonate-mediated
442 activation of murine and human sperm. *PLoS One* **5**, e15061 (2010).
- 443 31. Rebello, G. et al. Apoptosis-inducing signal sequence mutation in carbonic
444 anhydrase IV identified in patients with the RP17 form of retinitis pigmentosa. *Proc*
445 *Natl Acad Sci U S A* **101**, 6617-6622 (2004).
- 446 32. Yang, Z. et al. Mutant carbonic anhydrase 4 impairs pH regulation and causes retinal
447 photoreceptor degeneration. *Hum Mol Genet* **14**, 255-265 (2005).
- 448 33. Szklarczyk, D. et al. STRING v10: protein-protein interaction networks, integrated
449 over the tree of life. *Nucleic Acids Res* **43**, D447-452 (2015).
- 450 34. Cline, M.S. et al. Integration of biological networks and gene expression data using
451 Cytoscape. *Nat Protoc* **2**, 2366-2382 (2007).
- 452 35. Antal, B., Chessel, A. & Carazo Salas, R.E. Mineotaur: a tool for high-content
453 microscopy screen sharing and visual analytics. *Genome Biol* **16**, 283 (2015).
- 454 36. Adebayo, S. et al. PhenolImageShare: an image annotation and query infrastructure.
455 *Journal of biomedical semantics* **7**, 35 (2016).
- 456 37. Orlov, N. et al. WND-CHARM: Multi-purpose image classification using compound
457 image transforms. *Pattern Recognition Letters* **29**, 1684-1693 (2008).
- 458 38. Boulton, G., Rawlins, M., Vallance, P. & Walport, M. Science as a public enterprise:
459 the case for open data. *Lancet* **377**, 1633-1635 (2011).
- 460 39. Fuchs, F. et al. Clustering phenotype populations by genome-wide RNAi and
461 multiparametric imaging. *Molecular systems biology* **6**, 370 (2010).
- 462 40. Rohn, J.L. et al. Comparative RNAi screening identifies a conserved core metazoan
463 actinome by phenotype. *J Cell Biol* **194**, 789-805 (2011).
- 464 41. Breker, M., Gymrek, M. & Schuldiner, M. A novel single-cell screening platform
465 reveals proteome plasticity during yeast stress responses. *J Cell Biol* **200**, 839-850
466 (2013).
- 467 42. Thorpe, P.H., Alvaro, D., Lisby, M. & Rothstein, R. Bringing Rad52 foci into focus. *J*
468 *Cell Biol* **194**, 665-667 (2011).
- 469 43. Toret, C.P., D'Ambrosio, M.V., Vale, R.D., Simon, M.A. & Nelson, W.J. A genome-
470 wide screen identifies conserved protein hubs required for cadherin-mediated cell-
471 cell adhesion. *J Cell Biol* **204**, 265-279 (2014).
- 472 44. Fong, K.W. et al. Whole-genome screening identifies proteins localized to distinct
473 nuclear bodies. *J Cell Biol* **203**, 149-164 (2013).
- 474 45. Srikumar, T. et al. Global analysis of SUMO chain function reveals multiple roles in
475 chromatin regulation. *J Cell Biol* **201**, 145-163 (2013).
- 476 46. Doil, C. et al. RNF168 binds and amplifies ubiquitin conjugates on damaged
477 chromosomes to allow accumulation of repair proteins. *Cell* **136**, 435-446 (2009).

- 478 47. Karsenti, E. et al. A holistic approach to marine eco-systems biology. *PLoS Biol* **9**,
 479 e1001177 (2011).
- 480 48. Wawer, M.J. et al. Toward performance-diverse small-molecule libraries for cell-
 481 based phenotypic screening using multiplexed high-dimensional profiling. *Proc Natl*
 482 *Acad Sci U S A* **111**, 10911-10916 (2014).
- 483 49. Breinig, M., Klein, F.A., Huber, W. & Boutros, M. A chemical-genetic interaction map
 484 of small molecules using high-throughput imaging in cancer cells. *Molecular systems*
 485 *biology* **11**, 846 (2015).
- 486 50. Sero, J.E. et al. Cell shape and the microenvironment regulate nuclear translocation
 487 of NF-kappaB in breast epithelial and tumor cells. *Molecular systems biology* **11**, 790
 488 (2015).
- 489 51. Barr, A.R. & Bakal, C. A sensitised RNAi screen reveals a ch-TOG genetic
 490 interaction network required for spindle assembly. *Sci Rep* **5**, 10564 (2015).
- 491 52. Lawo, S., Hasegan, M., Gupta, G.D. & Pelletier, L. Subdiffraction imaging of
 492 centrosomes reveals higher-order organizational features of pericentriolar material.
 493 *Nat Cell Biol* **14**, 1148-1158 (2012).
- 494 53. Szymborska, A. et al. Nuclear pore scaffold structure analyzed by super-resolution
 495 microscopy and particle averaging. *Science* **341**, 655-658 (2013).
- 496 54. Dickerson, D. et al. High resolution imaging reveals heterogeneity in chromatin states
 497 between cells that is not inherited through cell division. *BMC Cell Biol* **17**, 33 (2016).
- 498 55. Pascual-Vargas, P. et al. RNAi screens for Rho GTPase regulators of cell shape and
 499 YAP/TAZ localisation in triple negative breast cancer. *Sci Data* **4**, 170018 (2017).
- 500 56. Yang, W. et al. Regulation of Meristem Morphogenesis by Cell Wall Synthases in
 501 Arabidopsis. *Curr Biol* **26**, 1404-1415 (2016).
- 502

503 Figure Legends

504

505 **Figure 1. Sampling of Phenotypes in the IDR.**

506 The numbers of samples per phenotype. Each sample represents a well from a micro-well
 507 plate in a screen or image from a dataset. Wells annotated as controls were not included.
 508 User submitted phenotype terms were mapped to the CMPO terms shown here. Colors
 509 represent higher-level groupings of phenotype terms. Point size shows the number of
 510 studies (group of related screens) each phenotype is linked to points of increasing size
 511 representing 1, 2, 3 or 4 studies respectively.

512

513

514 **Figure 2. Network Analysis of Genes Linked to the Elongated Cell Phenotype in the** 515 **IDR.**

516 A. Protein-protein interaction network based on the genes linked to the elongated cell
 517 phenotype (CMPO_0000077) in three distinct IDR studies. Genes from *S. pombe* (green,
 518 idr0001-A⁵), HeLa cell morphology (blue, idr0012-A³⁹) and HeLa Actinome (red, idr0008-B⁴⁰)
 519 are displayed with linkages (gray) from StringDB³³. To enable comparisons in Cytoscape,
 520 the human orthologues of *S. pombe* genes are used for the genes identified in idr0001-A.
 521 For more information, see Supplementary Note.

522

523 B. Zoomed view of network in A, with gene names. See Supplementary Note for the list of
 524 gene names used in the figure.

525

526 **Table 1. List of Datasets in IDR**

527 The phenotype column contains the number of submitted phenotypes. The number of genes, compounds or proteins identified as
 528 targets for analysis is listed in the Targets column and the 'Experiments' column lists the number of individual wells in HCS studies or
 529 imaging experiments in non-screen datasets.

530

Study	Species	Type	No. of screens or experiments	5D Images	Size (TB)	Phenotypes	Targets	Experiments	Reference
idr0001-graml-sysgro	<i>S. pombe</i>	gene deletion screen	1	109,728	10.06	19	3,005	18,432	5
idr0002-heriche-condensation	Human	RNAi screen	1	1,152	2.10	2	102	1,152	26
idr0003-breker-plasticity	<i>S. cerevisiae</i>	protein screen	1	97,920	0.20	14	6,234	32,640	41
idr0004-thorpe-rad52	<i>S. cerevisiae</i>	gene deletion screen	1	3,765	0.17	1	4,195	4,512	42
idr0005-toret-adhesion	<i>D. melanogaster</i>	RNAi screen	2	45,792	0.14	1	13,035	15,264	43
idr0006-fong-nuclearbodies	Human	protein localization screen	1	240,848	1.40	8	12,743	16,224	44
idr0007-srikumar-sumo	<i>S. cerevisiae</i>	protein localization screen	1	3,456	0.02	23	377	1,152	45
idr0008-rohn-actinome	<i>D. melanogaster</i> , Human	RNAi screen	2	55,944	0.12	46	12,826	26,496	40
idr0009-simpson-secretion	Human	RNAi screen	2	397,056	3.25	3	17,960	397,056	27
idr0010-doil-dnamage	Human	RNAi screen	1	56,832	0.08	2	18,675	56,832	46
idr0011-ledesmafernandez-dad4	<i>S. cerevisiae</i>	gene deletion screen	5	8,957	0.4	1	5209	8736	Under review

idr0012-fuchs-cellmorph	Human	RNAi screen	1	45,692	0.38	18	16,701	26,112	39
idr0013-neumann-mitocheck	Human	RNAi screen	2	200,995	14.54	18	18,393	206,592	4
idr0015-UNKNOWN-taraoceans	multi-species	geographic screen	1	32,776	2.49	0	84	84	47
idr0016-wawer-bioactivecompoundprofiling	Human	small molecule screen	1	869,820	3.19	2	29,542	144,000	48
idr0017-breinig-drugscreen	Human	small molecule screen	1	147,456	2.48	0	1,281	36,864	49
idr0018-neff-histopathology	<i>Mus musculus</i>	histopathology of gene knockouts	1	899	0.27	48	9	248	---
idr0019-sero-nfkappab	Human	HCS image analysis	1	25,872	0.03	0	198	2,156	50
idr0020-barr-chtog	Human	RNAi screen	1	36,960	0.03	2	241	1,232	51
idr0021-lawo-pericentriolarmaterial	Human	protein localization using 3D-SIM	1	414	0.0003	1	9	414	52
idr0023-szymborska-nuclearpore	Human	protein localization using dSTORM	1	524	0.0005	1	7	359	53
idr0027-dickerson-chromatin	<i>S. cerevisiae</i>	3D-tracking of tagged chromatin loci	1	229	0.03	0	8	112	54
idr0028-pascualvargas-rhogtpases	Human	RNAi screen	4	155,332	0.18	9	170	5544	55
idr0032-yang-meristem	<i>A. thaliana</i>	<i>in situ</i> hybridization	1	458	0.003	5	115	115	56
Sum			35	2,538,777	42	224	161,119	1,002,328	
Average				105,782	1.73	9	6,713	41,764	

531 Online Methods

532 **Architecture and Population of IDR**

533 IDR (<https://idr.openmicroscopy.org>) was built using open-source OMERO¹⁷ and Bio-Formats²⁴
534 as a foundation. Deployments are managed by Ansible playbooks along with re-usable roles on
535 an OpenStack-based cloud contained within the EMBL-EBI Embassy resource. Datasets (Table
536 1) were collected by shipped USB-drive or transferred by Aspera. Included datasets were
537 selected according to the criteria defined by the Euro-BioImaging/Elixir Data Strategy concept of
538 "reference images" ([http://www.eurobioimaging.eu/content-news/euro-bioimaging-elixir-image-](http://www.eurobioimaging.eu/content-news/euro-bioimaging-elixir-image-data-strategy)
539 [data-strategy](http://www.eurobioimaging.eu/content-news/euro-bioimaging-elixir-image-data-strategy)), which states that image datasets for publication should be related to published
540 studies, linked as much as possible to other resources and be candidates for re-use, re-
541 analysis, and/or integration with other studies.

542
543 Experimental and analytic metadata were submitted in either spreadsheets (CSV, XLS), PDF or
544 HDF5 files or a MySQL database, each using its own custom format. We converted these
545 custom formats to a consistent tabular format inspired by the MAGE and ISA-TAB
546 specifications^{21, 22} and combined into a single CSV file using a custom script (available in
547 <https://github.com/IDR/idr-metadata>) and imported into OMERO. Imaging metadata and binary
548 data were imported into OMERO using Bio-Formats. Experimental and analytic metadata were
549 stored using OMERO.tables, an HDF5-backed tabular data store used by OMERO. For each
550 dataset, metadata that were valuable for querying and search were copied to OMERO's key-
551 value-based Map Annotation facility²³. This means that different metadata types and elements
552 can be accessed using different parts of the OMERO API, depending on the search and
553 querying capabilities they require. For more information on the construction of queries, see
554 Supplementary Note.

555

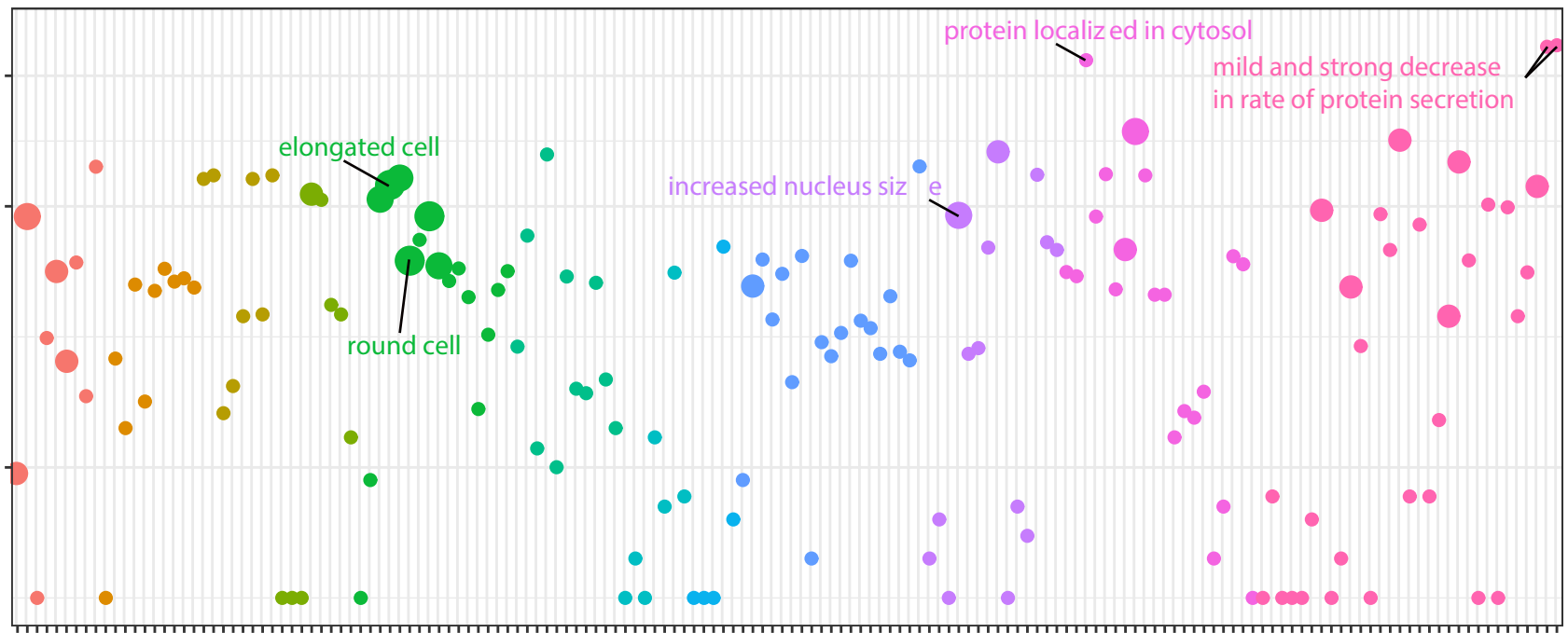
556 **Data Availability**

557 All datasets described in Table 1 and in this paper are available at
558 <https://idr.openmicroscopy.org>. All software for building and running the IDR and reading
559 metadata of the IDR datasets is open source and available at <https://github.com/IDR> and
560 <https://github.com/openmicroscopy>.

561

Number of samples

10000
1000
10



elongated cell

round cell

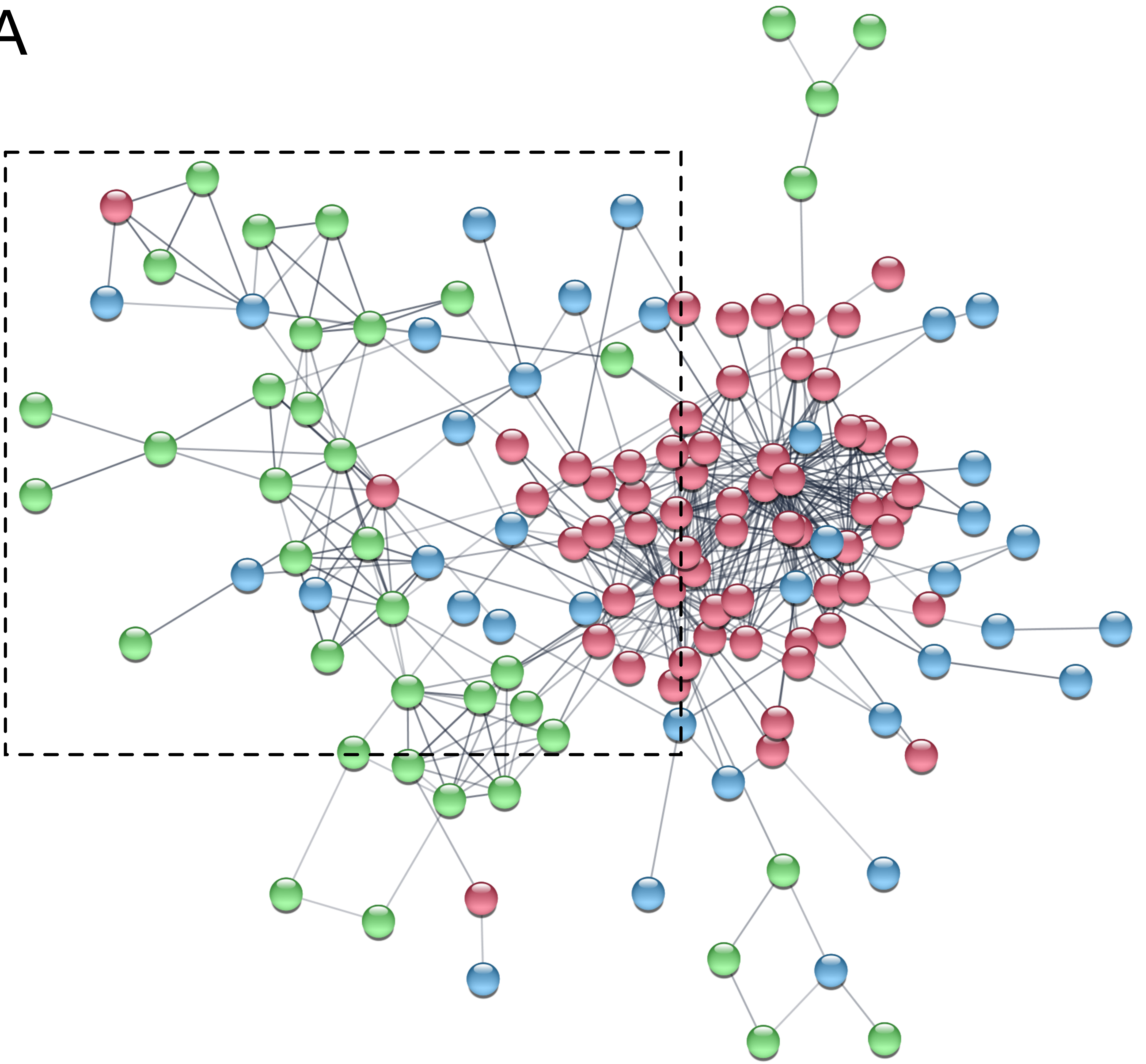
protein localized in cytosol

mild and strong decrease in rate of protein secretion

increased nucleus size

increased cell numbers
fewer cells with projections
more lamellipodia cells
more cells in M phase
more multinucleate cells
fewer aggregated cells in population
increased cell component number
increased number of microtubules
increased number of filopodia
increased amount of stress fibers
increased amount of ZIG-zag stress fibers
increased number of microtubule bundles
absence of mitotic chromosome decondensation
increased duration of mitotic prophase
mitotic metaphase plate congression
M phase arrested
M phase mitotic spindle
M phase delay
decreased duration of mitotic prophase
abnormal mitotic cell cycle phase
abnormal cell cycle
elongated cell
round cell
triangular shaped cell
geometric cell
5-shaped cell
curved cell
pear-shaped cell
stubby cell
fan-shaped cell
increased variability of cell shape in population
layered cells in population
more cells with metaphase microtubule spindles
fewer cells with interphase microtubule arrays
fewer cells with interphase microtubule arrays
more cells with G1 phase microtubule arrays
more cells with S phase microtubule arrays
increased cilium length
decreased cilium length
asymmetric lamellipodia
fan-shaped lamellipodia
cell component structure
increased cellular component
cell component morphology
increased number of actin filament
increased cortical actin cytoskeleton mass
disorganized cortical actin cytoskeleton
aggregated microtubules
microtubules nuclear
increased amount of punctate actin foci
increased amount of actin filaments
increased actin localized to the cytoplasm
shortened actin filaments
actin nuclear ring
shortened cytoplasmic microtubules
microtubule spindle morphology during apoptosis
abnormal microtubule cytoskeleton morphology during apoptosis
banded nucleus
increased nucleus size
decreased nucleus size
graped micronucleus
abnormal nucleus shape
bright nuclear body
phagocytic nuclear
protein localized in cytosol
protein localized in mitochondrion
protein localized in nuclear periphery
protein localized in punctate foci
protein localized in nucleus
protein localized in v. acicular membrane
protein localized in Cajal body
protein localized in paraspeckle
protein localized in PML body
protein localized in Golgi body
protein localized in centrosome
protein localized in centrosome
absence of protein localized in bud neck
absence of mitotic process
cell-matrix adhesion
absence of cell spreading
increased thickness of dendritic branching
cell apoptosis
missense mutation in DNA topoisomerase
abnormal cell growth
increased cell movement distance
increased cell movement velocity
loss of cell motility
aggregated cells in population
cell response to DNA damage
abnormal chromosome segregation
increased microtubule metabolic process
regulation of metastable kinetochore
negative regulation of protein import into nucleus
positive regulation of protein import into nucleus
cell cycle arrest
increased rate of protein secretion
mild decrease in rate of protein secretion
strong decrease in rate of protein secretion

A



B

