

Cell Painting Gallery: an open resource for image-based profiling



Image-based or morphological profiling is a rapidly expanding field wherein cells are ‘profiled’ by extracting hundreds to thousands of unbiased, quantitative features from images of cells that have been genetically or chemically perturbed. It is the least expensive high-dimensional profiling technique to date and offers single-cell resolution. The Cell Painting assay is the most popular imaged-based profiling assay; in it, six small-molecule dyes label eight cellular compartments and are imaged in five fluorescence channels (plus brightfield, if desired). Images are segmented and thousands of measurements are then made within the nucleus, cytoplasm and whole cell, describing quantitative traits such as size, shape, stain intensity and stain texture. First published in 2013 (ref. 1), the standard protocol was updated in 2016 and more recently in 2023 (ref. 2).

The field of bioimaging, like most scientific fields, is transitioning toward making data increasingly FAIR (findable, accessible, interoperable and reusable)³. In this vein, we have created the Cell Painting Gallery, a publicly available collection of Cell Painting datasets with granular dataset descriptions and comprehensive documentation, including access instructions, at <https://github.com/broadinstitute/cellpainting-gallery>. It is hosted by Amazon Web Services (AWS) on the Registry of Open Data, and anyone can access the data without requiring an AWS account. As of May 2024, the Cell Painting Gallery holds 688 TB of image and associated numerical data. It includes the largest publicly available Cell Painting dataset, in terms of perturbations tested (Joint Undertaking for Morphological Profiling, or JUMP), along with many other canonical datasets using Cell Painting and close derivatives (Table 1). Other sources of publicly available Cell Painting datasets include Recursion (<https://RxRx.ai>) and Image Data Resource (IDR; <https://idr.openmicroscopy.org>).

Cell Painting has been widely used for varied biological applications, including characterizing gene function (through

overexpression, knockout, knockdown and variant characterization), drug toxicity prediction, mechanism of action identification, and multiple other stages of drug discovery (summarized in ref. 4). In service of these applications, deep learning methods have been developed that improve on traditional machine learning methods across diverse aspects of high-content imaging, including image quality, object segmentation, object and image feature extraction, stain prediction, and object and image classification⁵. Curating large, diverse Cell Painting datasets into the Cell Painting Gallery enhances the potential for both data mining for biological discovery and the development of new machine and deep learning techniques, particularly representation learning and multimodal learning. To that end, we have implemented specific data and metadata guidelines, provided comprehensive download instructions, reprocessed old datasets, converted datasets to next-generation file formats, and worked with external organizations to make the Cell Painting Gallery browsable with their infrastructure, all detailed below.

Datasets in Cell Painting Gallery follow specific data and metadata organizational requirements, an important trait of FAIR data. This allows images and numerical data to easily be downloaded or accessed separately, simplifying application-specific data download. Our data validator analyzes the contents of the gallery and generates customizable reports of attributes such as data integrity and dataset completion.

To make the Cell Painting Gallery accessible to all researchers regardless of computational or cloud expertise, we provide detailed documentation with instructions for accessing and downloading data from AWS, including curating varied example workflows that access data in the gallery.

The field of morphological profiling has undergone major developments in the last 10 years, and, as in any scientific field, data that were state-of-the-art at publication may be a challenge to use 10 years later. To bring important historical datasets up to date, we have

reprocessed several key historical datasets – cpg0012-waver-bioactivecompoundprofiling, cpg0017-rohban-pathways and cpg0031-caicedo-cmvip – using state-of-the-art informatics² so that these modernized datasets now have the same feature set as the -136,000 chemical and genetic perturbations in cpg0016-jump, the JUMP Cell Painting dataset.

Proprietary image file formats hinder the reuse of both images and workflows. To address this, we have begun converting datasets, such as cpg0004-lincs, to OME-Zarr, a next-generation file format optimized for data FAIRness and cloud-based usage^{6,7}. These conversions, performed with tools like Distributed-OMEZarrCreator⁸, also improve interoperability with IDR, soon allowing Cell Painting Gallery data to be browsed through IDR.

The Cell Painting Gallery is continually expanding. Though the majority of the data, apart from JUMP, have been collected at the Broad Institute, we welcome data from other sources and detail in our documentation how to nominate datasets for inclusion. We also welcome community contributions to our documentation, including example workflows and publications that access data in the Gallery. We intend to maintain the gallery beyond the duration of any individual program or grant.

Data availability

All data are freely available in the Cell Painting Gallery, hosted by AWS on the Registry of Open Data. An AWS account is not necessary for accessing data. Documentation is provided at <https://broadinstitute.github.io/cellpainting-gallery>.

Code availability

The Cell Painting Gallery landing page is at <https://github.com/broadinstitute/cellpainting-gallery>. Cell Painting Gallery documentation is at <https://broadinstitute.github.io/cellpainting-gallery>. The Cell Painting Gallery data validator and other tools that interact with Cell Painting Gallery are available at <https://github.com/broadinstitute/cpg>.

Correspondence










Table 1 | Complete datasets available in the Cell Painting Gallery as of publication

Dataset name	Description	Publication to cite	Objects	Total data size	Image data size	Numerical data size
cpg0000-jump-pilot	300+ compounds and 160+ genes (CRISPR knockout and overexpression) profiled in A549 and U2OS cells at two time points	Chandrasekaran, S. N. et al. Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. Preprint at <i>bioRxiv</i> https://doi.org/10.1101/2022.01.05.475090 (2022)	5.2 M	12.3 TB	6.1 TB	6.1 TB
cpg0001-cellpainting-protocol	300+ compounds profiled in U2OS cells using several different modifications of the Cell Painting protocol; description of Cell Painting v3	Cimini, B. A. et al. Optimizing the Cell Painting assay for image-based profiling. <i>Nat. Protoc.</i> https://doi.org/10.1038/s41596-023-00840-9 (2023)	9.6 M	40.3 TB	18.7 TB	21.6 TB
cpg0002-jump-scope	300+ compounds profiled in U2OS using different microscopes and settings	Tromans-Coia, C. et al. Assessing the performance of the Cell Painting assay across different imaging systems. <i>Cytometry A</i> 103 , 915–926 (2023)	2.6 M	16.7 TB	12.5 TB	4.2 TB
cpg0003-rosetta	28,000+ genes and compounds profiled in Cell Painting and L1000 gene expression	Haghighi, M., Caicedo, J. C., Cimini, B. A., Carpenter, A. E. & Singh, S. High-dimensional gene expression and morphology profiles of cells across 28,000 genetic and chemical perturbations. <i>Nat. Methods</i> 19 , 1550–1557 (2022)	51	8.5 GB	0	8.5 GB
cpg0004-lincs	1,571 compounds across 6 doses in A549 cells	Way, G. P. et al. Morphology and gene expression profiling provide complementary information for mapping cell state. <i>Cell Syst.</i> 13 , 911–923.e9 (2022)	70.5 M	65.7 TB	61.9 TB	3.8 TB
cpg0010-caie-drugrespons	MCF-7 breast cancer cells treated with 113 small molecules at eight concentrations	Caie, P. D. et al. High-content phenotypic profiling of drug response signatures across distinct cancer cells. <i>Mol. Cancer Ther.</i> 9 , 1913–1926 (2010)	1.1 M	239.2 GB	98.4 GB	140.8 GB
cpg0011-lipocyteprofiler	Variety of lipocytes in different metabolic states and with genetic and drug perturbations	Laber, S. et al. Discovering cellular programs of intrinsic and extrinsic drivers of metabolic traits using LipocyteProfiler. <i>Cell Genomics</i> 3 , 100346 (2023)	143 K	1.2 TB	1.2 TB	16 MB
cpg0012-wawer-bioactivecompoundprofiling	30,000-compound dataset in U2OS cells	Wawer, M. J. et al. Toward performance-diverse small-molecule libraries for cell-based phenotypic screening using multiplexed high-dimensional profiling. <i>Proc. Natl Acad. Sci. USA</i> 111 , 10911–10916 (2014) Bray, M.-A. et al. A dataset of images and morphological profiles of 30 000 small-molecule treatments using the Cell Painting assay. <i>Gigascience</i> 6 , 1–5 (2017)	11 M	10.7 TB	3.1 TB	7.6 TB
cpg0015-heterogeneity	2,200+ compounds and 200+ genes profiled in U2OS cells	Rohban, M. H., Abbasi, H. S., Singh, S. & Carpenter, A. E. Capturing single-cell heterogeneity via data fusion improves image-based profiling. <i>Nat. Commun.</i> 10 , 2082 (2019)	619	204 GB	0	204 GB
cpg0016-jump	116,000+ compounds and 16,000+ genes (CRISPR knockout and overexpression) profiled in over 1.5 billion U2OS cells	Chandrasekaran, S. N. et al. JUMP Cell Painting dataset: morphological impact of 136,000 chemical and genetic perturbations. Preprint at <i>bioRxiv</i> https://doi.org/10.1101/2023.03.23.534023 (2023)	115.3 M	358.4 TB	115.3 TB	243 TB
cpg0017-rohban-pathways	323 genes overexpressed in U2OS cells; original images re-profiled in 2023	Rohban, M. H. et al. Systematic morphological profiling of human gene and allele function via Cell Painting. <i>Elife</i> 6 , (2017)	305 K	321 GB	189 GB	132 GB
cpg0018-singh-seedseq	U2OS cells treated with each of 315 unique short hairpin RNA sequences	Singh, S. et al. Morphological profiles of RNAi-induced gene knockdown are highly reproducible but dominated by seed effects. <i>PLoS One</i> 10 , e0131370 (2015)	138 K	247.1 GB	247.1 GB	0

Table 1 (Continued) | Complete datasets available in the Cell Painting Gallery as of publication

Dataset name	Description	Publication to cite	Objects	Total data size	Image data size	Numerical data size
cpg0019-moshkov-deepprofiler	8.3 million single cells from 232 plates, across 488 treatments from 5 public datasets, used for learning representations	Moshkov, N. et al. Learning representations for image-based profiling of perturbations. <i>Nat. Commun.</i> 15 , 1594 (2024)	9.3 M	522 GB	482 GB	40 GB
cpg0021-periscope	30 million cells with 20,000 single-gene knockouts in pooled format; A549 cells and HeLa cells in two growth media	Ramezani, M. et al. A genome-wide atlas of human cell morphology. Preprint at <i>bioRxiv</i> https://doi.org/10.1101/2023.08.06.552164 (2023)	7.1 M	56.0 TB	45.0 TB	11.0 TB
cpg0022-cmqtl	297 induced pluripotent stem cell lines	Tegtmeyer, M. et al. High-dimensional phenotyping to define the genetic basis of cellular morphology. <i>Nat. Commun.</i> 15 , 347 (2024)	702 K	3.7 TB	2.8 TB	945 GB
cpg0028-kelley-resistance	Bortezomib-resistant HCT116 clones	Kelley, M. E. et al. High-content microscopy reveals a morphological signature of bortezomib resistance. <i>Elife</i> 12 , (2023)	1 M	4.1 TB	1.9 TB	2.2 TB
cpg0030-gustafsdottir-cellpainting	U2OS cells treated with each of 1,600 known bioactive compounds; description of Cell Painting v1	Gustafsdottir, S. M. et al. Multiplex cytological profiling assay to measure diverse cellular states. <i>PLoS One</i> 8 , e80999 (2013)	346 K	234 GB	234 GB	0.3 GB
cpg0031-caicedo-cmvip	ORF overexpression of 596 alleles of 53 genes in A549 cells	Caicedo, J. C. et al. Cell Painting predicts impact of lung cancer variants. <i>Mol. Biol. Cell</i> 33 , ar49 (2022)	553 K	802 GB	605 GB	197 GB

Total data size (complete and in-progress datasets) as of May 2024 is 688 TB. K, thousand; M, million.

Erin Weisbart  , **Ankur Kumar** , **John Arevalo** , **Anne E. Carpenter** , **Beth A. Cimini**  & **Shantanu Singh**  
Imaging Platform Department, Broad Institute of MIT and Harvard, Cambridge, MA, USA.
 e-mail: eweisbar@broadinstitute.org; shantanu@broadinstitute.org

Published online: 02 September 2024

References

- Gustafsdottir, S. M. et al. *PLoS One* **8**, e80999 (2013).
- Cimini, B. A. et al. *Nat. Protoc.* **18**, 1981–2013 (2023).
- Schmied, C. et al. *Nat. Methods* **21**, 170–181 (2024).
- Chandrasekaran, S. N., Ceulemans, H., Boyd, J. D. & Carpenter, A. E. *Nat. Rev. Drug Discov.* **20**, 145–159 (2021).
- Carreras-Puigvert, J. & Spjuth, O. *Curr. Opin. Struct. Biol.* **87**, 102842 (2024).

- Moore, J. et al. *Nat. Methods* **18**, 1496–1498 (2021).
- Moore, J. et al. *Histochem. Cell Biol.* **160**, 223–251 (2023).
- Weisbart, E. & Cimini, B. A. *Nat. Methods* **20**, 1120–1121 (2023).

Acknowledgements

We thank E. Chu and N. Ragusa of AWS for assistance in getting the Cell Painting Gallery set up, along with members of the Cimini and Carpenter-Singh labs for their feedback on this project and manuscript. We also thank T. Treis and F. Theis for publicly sharing their workflow accessing gallery data. This study was supported by Calico Life Sciences LLC; Merck Sharp & Dohme LLC, a subsidiary of Merck & Co., Inc.; NIH P41 GM135019 (to B.A.C. and A.E.C.), NIH R35 GM122547 (to A.E.C.); and CZI grant DAF2020-225720 (grant <https://doi.org/10.37921/977328pjbvca>) from the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation (funder <https://doi.org/10.13039/100014989>) (to B.A.C.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

S.S., B.A.C. and A.E.C. conceived and supervised the project and revised the manuscript. A.K. and J.A. wrote data validation scripts and revised the manuscript. E.W. curated the data, wrote documentation and wrote the manuscript.

Competing interests

S.S. and A.E.C. serve as scientific advisors for companies that use image-based profiling and Cell Painting (A.E.C.: Recursion, SyzOnc, Quiver Bioscience, S.S.: Waypoint Bio, Dewpoint Therapeutics, Deepcell) and receive honoraria for occasional talks at pharmaceutical and biotechnology companies. The other authors declare no competing interests.

Additional information

Peer review information *Nature Methods* thanks Isabel Kemmer, Aastha Mathur, Ugis Sarkans and the other, anonymous reviewer(s) for their contribution to the peer review of this work.