Title

JUMP Cell Painting dataset: morphological impact of 136,000 chemical and genetic perturbations

Authors

Srinivas Niranj Chandrasekaran[1], Jeanelle Ackerman[1], Eric Alix[2], D. Michael Ando[3], John Arevalo[1], Melissa Bennion[4], Nicolas Boisseau[5], Adriana Borowa[6], Justin D. Boyd[7], Laurent Brino[8], Patrick J. Byrne[1], Hugo Ceulemans[9], Carolyn Ch'ng[10], Beth A. Cimini[1], Djork-Arne Clevert[11], Nicole Deflaux[12], John G Doench[1], Thierry Dorval[5], Regis Doyonnas[13], Vincenza Dragone[14], Ola Engkvist[15], Patrick W. Faloon[4], Briana Fritchman[1], Florian Fuchs[16], Sakshi Garg[14], Tamara J. Gilbert[7], David Glazer[12], David Gnutt[17], Amy Goodale[1], Jeremy Grignard[5], Judith Guenther[11], Yu Han[1], Zahra Hanifehlou[8], Santosh Hariharan[13], Desiree Hernandez[1], Shane R Horman[18], Gisela Hormel[14], Michael Huntley[19], Ilknur Icke[20], Makiyo Iida[21], Christina B. Jacob[8], Steffen Jaensch[9], Jawahar Khetan[19], Maria Kost-Alimova[1], Tomasz Krawiec[6], Daniel Kuhn[14], Charles-Hugues Lardeau[9], Amanda Lembke[19], Francis Lin[19], Kevin D. Little[4], Kenneth R. Lofstrom[4], Sofia Lotfi[5], David J. Logan[7], Yi Luo[18], Franck Madoux[19], Paula A. Marin Zapata[11], Brittany A. Marion[1], Glynn Martin[2], Nicola Jane McCarthy[2], Lewis Mervin[22], Lisa Miller[1], Haseeb Mohamed[19], Tiziana Monteverde[22], Elizabeth Mouchet[22], Barbara Nicke[11], Arnaud Ogier[8], Anne-Laure Ong[5], Marc Osterland[11], Magdalena Otrocka[6], Pieter J. Peeters[9], James Pilling[22], Stefan Prechtl[11], Chen Qian[22], Krzysztof Rataj[6], David E Root[1], Sylvie K. Sakata[23], Simon Scrace[2], Hajime Shimizu[21], David Simon[19], Peter Sommer[8], Craig Spruiell[19], Iffat Sumia[5], Susanne E Swalley[4], Hiroki Terauchi[21], Amandine Thibaudeau[8], Amy Unruh[12], Jelle Van de Waeter[24], Michiel Van Dyck[9], Carlo van Staden[19], Michał Warchoł[6], Erin Weisbart[1], Amélie Weiss[8], Nicolas Wiest-Daessle[8], Guy Williams[22], Shan Yu[18], Bolek Zapiec[14], Marek Żyła[6], Shantanu Singh[1#], Anne E. Carpenter[1#]

co-corresponding author

Affiliations

1. Broad Institute, Cambridge, MA, USA
2. Horizon Discovery Ltd., Cambridge, Cambridgeshire, UK
3. Google Applied Sciences; Mountain View, CA, USA
4. Biogen, Cambridge, MA, USA
5. Institut de Recherche Servier, Paris-Saclay, France
6. Ardigen S.A, Kraków, Poland
7. Pfizer Worldwide Research, Development and Medical, Cambridge, MA, USA
8. Ksilink, Strasbourg, France
9. Janssen Pharmaceutica N.V., Beerse, Belgium
10. Amgen, Burnaby, BC, Canada
11. Bayer AG, Research & Development, Pharmaceuticals, Berlin, BN, Germany
12. Verily Life Sciences, South San Francisco, CA, USA
13. Pfizer Worldwide Research, Development and Medical, Groton, CT, USA
14. Merck Healthcare KGaA, Darmstadt, Germany
15. Discovery Sciences, R&D, AstraZeneca, Gothenburg, Sweden & Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg, Sweden
16. NUVISAN ICB GmbH, Lead Discovery Screening, Berlin, BN, Germany
17. Bayer AG, Research & Development, Pharmaceuticals, Wuppertal, NRW, Germany

18. Takeda Development Center Americas Inc., San Diego, CA, USA
19. Amgen, Thousand Oaks, CA, USA
20. Takeda Development Center Americas Inc., Cambridge, MA, USA
21. Eisai, Tsukuba, Japan
22. AstraZeneca BioPharmaceuticals R&D, Cambridge, UK
23. Pfizer Worldwide Research, Development and Medical, La Jolla, San Diego, CA, USA
24. Charles River Laboratories, Beerse, Belgium

Abstract

Image-based profiling has emerged as a powerful technology for various steps in basic biological and pharmaceutical discovery, but the community has lacked a large, public reference set of data from chemical and genetic perturbations. Here we present data generated by the Joint Undertaking for Morphological Profiling (JUMP)-Cell Painting Consortium, a collaboration between 10 pharmaceutical companies, six supporting technology companies, and two non-profit partners. When completed, the dataset will contain images and profiles from the Cell Painting assay for over 116,750 unique compounds, over-expression of 12,602 genes, and knockout of 7,975 genes using CRISPR-Cas9, all in human osteosarcoma cells (U2OS). The dataset is estimated to be 115 TB in size and capturing 1.6 billion cells and their single-cell profiles. File quality control and upload is underway and will be completed over the coming months at the Cell Painting Gallery: https://registry.opendata.aws/cellpainting-gallery. A portal to visualize a subset of the data is available at https://phenaid.ardigen.com/jumpcpexplorer/.

Introduction

The pharmaceutical industry needs disruptive technologies to rapidly reduce the cost and failure rates for getting life-changing medicines to patients. The image-based microscopy profiling assay, Cell Painting [1], has shown promise in several steps in the drug discovery pipeline [2], including disease phenotype identification [3], hit identification [4,5] assay activity prediction [6], toxicity detection [7,8] and mechanism of action determination [9], as well as in basic biological research such as functional genomics [10]. In this assay, eight cellular components are stained with six inexpensive dyes and imaged in five channels on a fluorescence microscope. Image analysis software identifies individual cells and extracts a few thousand features from each, producing single-cell profiles. The technology is significantly less expensive and higher-throughput than transcriptomic and proteomic profiling. Cell images contain a vast amount of very quantifiable information about the status of the cell: for example, whether it is diseased, whether it is responding to a drug treatment, and whether a pathway has been disrupted by a genetic mutation.

Just as is the case for genomics and transcriptomics, a public reference database is critical for image-based profiling. For many applications, a query sample's image-based profile can be matched to similar (or opposite) profiles in the reference database, yielding hypotheses about the sample of interest; this guilt-by-similarity strategy is the basis of matching query compounds to annotated compounds to discern a mechanism of action, or to identify potential regulators of a given gene's pathway by matching the gene's profile to candidate compounds [10]. A large structured reference database can also be used to train machine learning models to predict compounds' assay activity [6] to identify promising compounds to test physically [4,5]. Image data can also be used in representation learning, to teach deep learning models biologically useful embeddings [11,12].

Large, general image repositories have launched over the past decade [13,14], and offer a tremendous benefit to the scientific community. Nevertheless, their use for many applications in image-based profiling has been limited because few public image sets in such repositories use a standardized assay like Cell Painting. The largest public Cell Painting datasets to date have been produced at (a) the Broad Institute, with 30,616 compounds tested [15] and a few hundred genetic perturbations [10,16], and (b) Recursion, with several datasets of roughly a thousand perturbations each (RNA interference, immune stimulants [4], and compounds, often in dose-response and/or multiple cell types, https://www.rxrx.ai). These large standardized datasets have proven powerful in many applications, including identifying candidate therapeutics and targets. They have demonstrated the value and provided the motivation to create a much larger public dataset, while also raising the question whether large datasets created at multiple experimental laboratories could be productively combined.

Here, we present the largest, freely available Cell Painting dataset of chemical and genetic perturbations, systematically acquired across 12 pharmaceutical and academic partner sites in five replicates, with the support of six technology companies. Although the data is in the process of final quality control and upload, we wanted to provide the community with sufficient information to design experiments and analyses, and to begin using the data as more of it becomes publicly available over the next few months.

Results

Data production and organization

The Joint Undertaking for Morphological Profiling (JUMP)-Cell Painting Consortium is made of 10 pharmaceutical/biotechnology companies, two non-profit organizations and six supporting partners (https://jump-cellpainting.broadinstitute.org/). The goal of the consortium was to create the largest publicly available Cell Painting dataset and to recommend best practices for running Cell Painting experiments and profile generation, while simultaneously developing methods for community-oriented data analysis and storage.

The JUMP dataset comprises four subsets: a large production dataset of cells perturbed by three different perturbation modalities–chemical compounds (small molecules), overexpression of genes using Open Reading Frames (ORFs) and gene knockout by Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) guides (cpg0016), and three pilot datasets in which cells were perturbed under various experimental and imaging conditions (cpg0000 [17], cpg0001 [1] and cpg0002 [18]; note that the first two of these were previously released). We use the term “knockout” here because the majority of resulting alleles have frameshift mutations, preventing further protein production; however, varying levels of proteins may remain at the time of analysis, and some cells will not contain knockout alleles at all. Across all the datasets, there are more than 116,000 chemical perturbations and 15,000 genes perturbed (by ORF, CRISPR, or both). The size and overlap between these datasets are shown in Figure 1.

The primary JUMP dataset (cpg0016), apart from being the largest, has two main features that make it ideal as a reference dataset: 1) the compound dataset in cpg0016 was generated across 11 data producing sites (or “sources”). Each source, apart from source_7, exchanged their nominated compounds with either two or four other sources where they were assayed using different instruments and microscopes (details are provided in the Methods section). This technical variability makes the data more robust as a reference dataset. 2) JUMP-Target-2-Compound, a positive control plate of 306 diverse compounds, was run with every batch of

data generation. These plates not only allow alignment of data within the JUMP dataset, but also with future datasets generated outside the consortium.

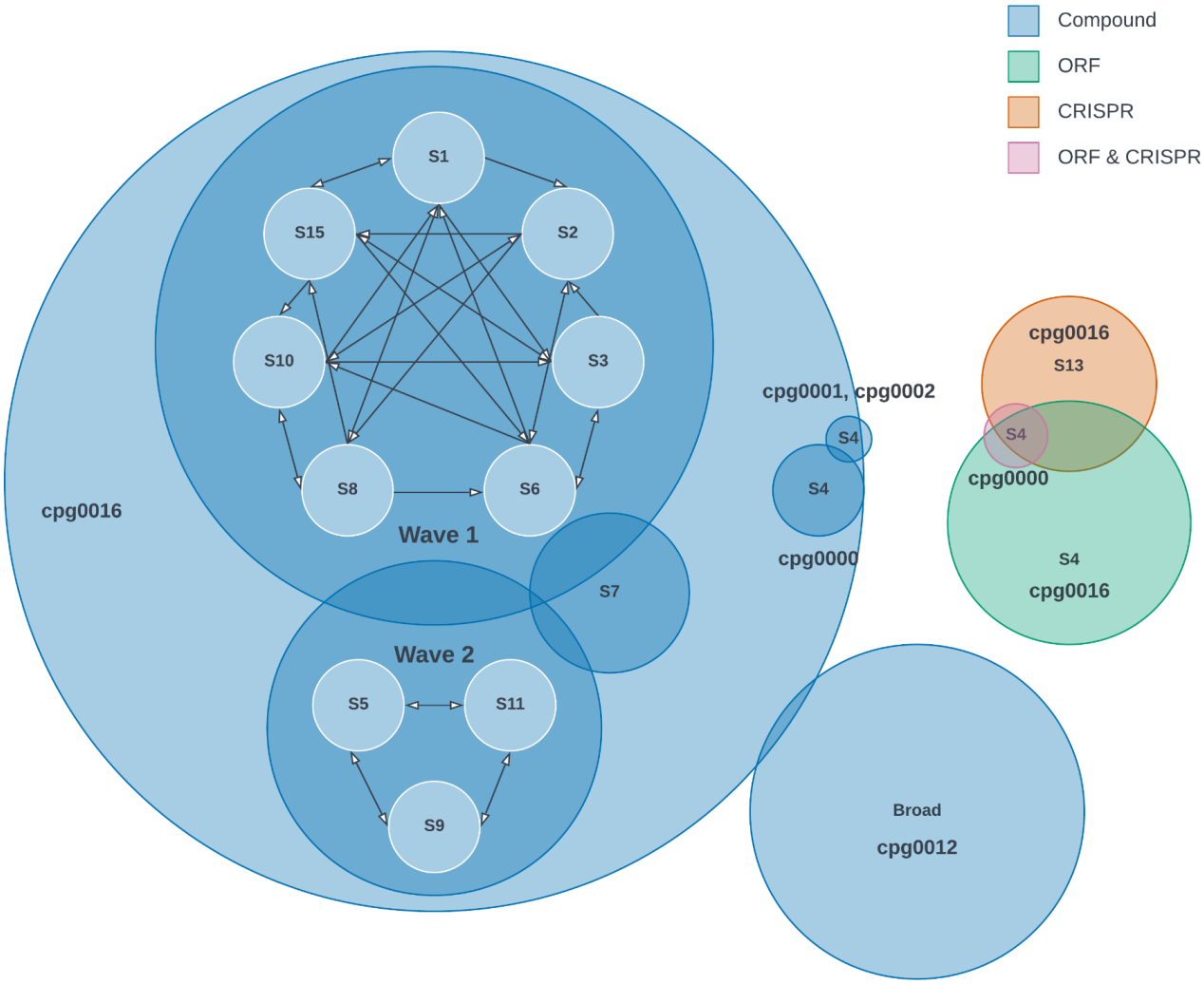

Figure 1: **Data and perturbations in the JUMP-Cell Painting dataset**. The JUMP Cell Painting dataset is a collection of four datasets, namely, cpg0000, cpg0001, cpg0002, and cpg0016 (these datasets are described in the Methods section). All datasets were generated at the Broad (known as source_4, or S4), except cpg0016, which was generated across the 12 partner sites. Each source number corresponds to a different site, except S7 and S13, which correspond to the same site. The genetic perturbation data was generated at individual sites (CRISPR at S13 and ORFs at S4) while typically the five replicates of each compound were generated at two to four sites besides the site that nominated the compound. The arrows denote the compound exchange logistics (Supplementary Table 3). There are overlaps in compounds or genes tested in the different datasets, but the size of the circles and their overlap are schematic and thus not precisely to scale. Overlap of the JUMP-Cell Painting dataset is shown with cpg0012 [15]), an existing public dataset of 30,616 bioactive compounds that was previously generated at the Broad and reprocessed using the JUMP processing pipeline (as described in the methods section).

Visualizing ORF profiles

Of the three perturbation modalities–compound, CRISPR-guides and ORF–the ORF dataset is currently complete and released publicly (all five replicates of each perturbation have been released). The two-dimensional UMAP [19] representations of the normalized, feature-selected ORF profiles (steps to generate these profiles are described in the methods section) are shown in Figure 2. We find that the control wells do not cluster separately from the ORF treatment wells (Figure 2a, Supplementary Figure 1). This could be due to several reasons: 1) ORF treatments are generally indistinguishable from controls, 2) the negative and positive controls are not optimally fulfilling their role (see Methods details – optimal negative controls are difficult to define, as any non-human ORF may not be completely inert, and the proteins we chose are known to have an impact on human cells) or 3) technical effects, such as batch effect and well position effect, are stronger than the ORF induced effects on cellular morphology. To disentangle the ORF induced biological effects from the technical effects, we plotted additional views of the UMAP representation. We do not observe significant batch effects when the UMAP is colored by the experimental batch (Figure 2b, Supplementary Figure 4), perhaps because the entire ORF dataset was generated at a single site (source_4). But, when we color the UMAP by row name (Figure 2c, Supplementary Figures 2 and 5) and column name (Figure 2d, Supplementary Figures 3 and 6), we observe a clear striping pattern as wells from the same row or column cluster together. We observe this pattern irrespective of the perturbation type (Supplementary Figures 2 and 3) and experimental batch (Supplementary Figures 5 and 6); the pattern cannot be explained by particular samples being consistently in those rows/columns. Correcting for the well position effects may help recover true biological signals in this dataset, which is particularly important because all the replicates for each ORF treatment occupy the same well position across plates, an unfortunate practical limitation of plating the samples.

We created a portal to visualize a subset of the data at https://phenaid.ardigen.com/jumpcpexplorer/ and a third-party software company has created an additional portal at https://www.springdiscovery.com/jump-cp.

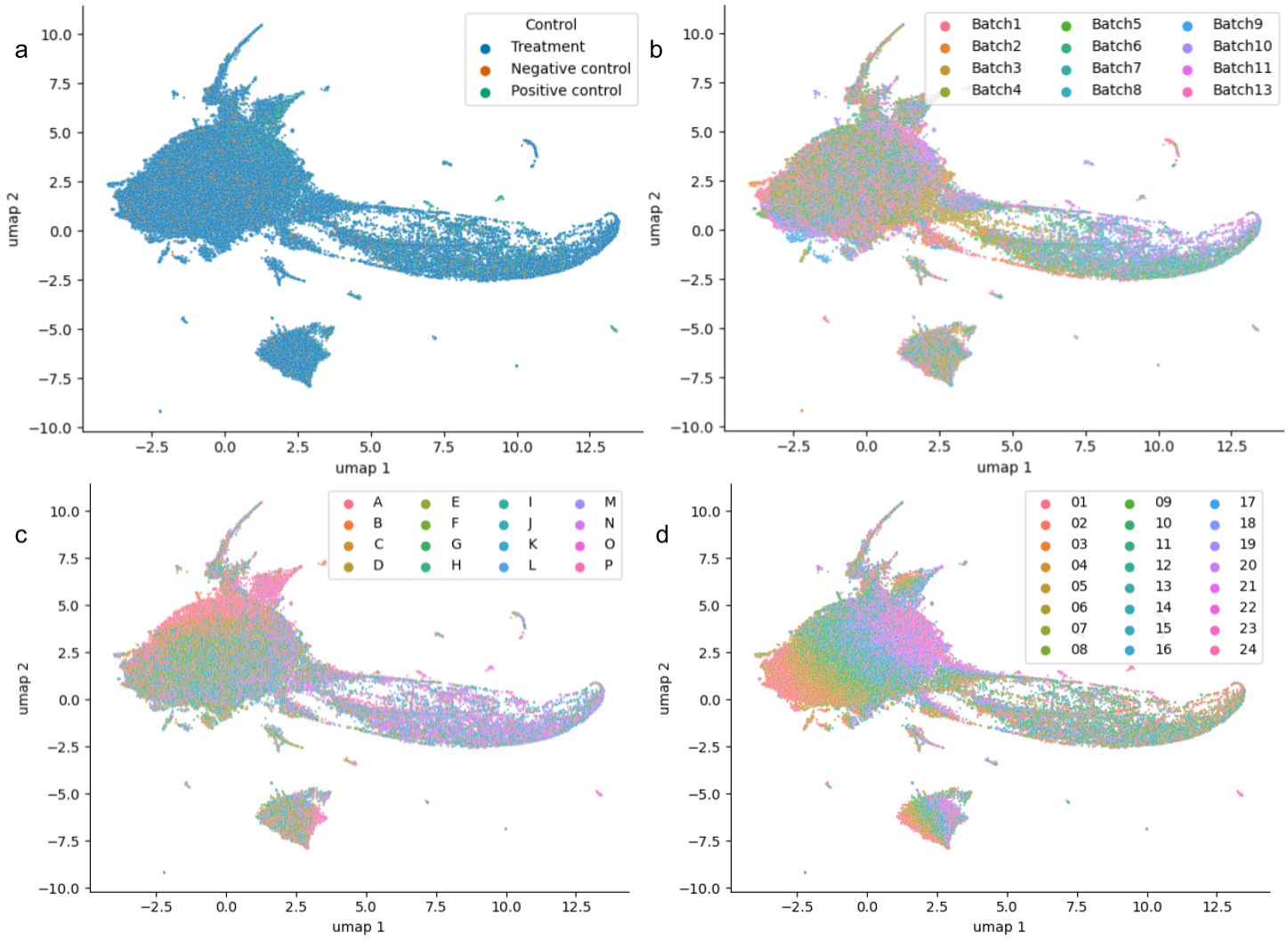

Figure 2: **UMAP of ORF samples**. All the wells in the ORF perturbed plates are shown in this UMAP, which are colored by treatment type in a), by batch in b), by row name in c) and column name in d). The entire ORF dataset was generated at source_4. Additional views of the UMAP representations are shown in Supplementary Figures 1-6.

Discussion

Although the full dataset is not yet available online, we have described our experimental design and methodology to assist the community in designing experiments and analyses and to begin using the data. For our own part, we have many plans for using the data, for example, in applications summarized in the introduction.

We also plan to create and share improved versions of the data as follows. We will extract profiles from the images using deep learning models, both an ImageNet 21k pretrained model (using the EfficientNetV2-S architecture) [20] and one learned specifically for Cell Painting images [11]. We intend to carry out deep-learning based quality control to filter out images that are blurry or containing debris. We will assess alternatives [21] and carry out batch correction to better align the data within batches at each data collection site and across sites, as well as to correct the well-position effects (also known as plate-layout effects) where row and column position impacts the morphological profiles. We will assess methods for combining single-cell data into profiles

for a sample, as the current approach is based on simple population means. We will explore better methods for matching samples, particularly genes, to compounds (https://github.com/jump-cellpainting/genemod). We will assess whether destaining Cell Painted gene overexpression samples and restaining to detect a tag on the desired protein can be used to create improved profiles from the subset of cells expressing an appropriate level of protein. We will extract profiles from the unlabeled brightfield images to assess their information content. We will apply multivariate analysis to analyze the extracted features, including use of soft classifiers on a subset of samples. We will assess whether Cell Painting image-based data can be effectively integrated with secretome data captured using nELISA®, a protein profiling platform offered by Nomic Bio.

A world of opportunities is now open with this large perturbational dataset that contains both chemical and genetic perturbations. We invite the scientific community to make maximal use of the JUMP-Cell Painting dataset and to ask further clarifying questions via GitHub issues at https://github.com/jump-cellpainting/datasets/issues.

Methods

The JUMP Cell Painting dataset is a collection of several datasets that were either generated or reprocessed by the JUMP Cell Painting Consortium. The primary dataset (cpg0016; production data) was generated during the data production phase of the JUMP-CP project. It comprises a compound dataset (compound production data), an Open Reading Frame (ORF) overexpression dataset (ORF production data) and a CRISPR knockout dataset (CRISPR production data). In addition, three datasets were generated during the pilot phase of the project (pilot data). These include a previously released compound dataset to test different experimental/staining conditions (cpg0001; CPPilot data [1]), another previously released dataset to test different perturbation modalities (cpg0000; CPJUMP1 data [17]) and a compound dataset to test different microscopes (cpg0002; CPJUMP-Scope1 data [18]). In the following sections, unless specified explicitly, all the details about experimental conditions and analysis pertain to the cpg0016 data. Details about the other datasets are described in other manuscripts describing those datasets.

Chemical and genetic perturbations

The JUMP Cell Painting dataset consists of cells that were perturbed by both chemical (small molecules, also called compounds) and genetic (overexpression using ORF and knockout using CRISPR-Cas9) perturbations.

Chemical perturbation selection

For generating the production data, the JUMP consortium partners nominated and provided 116,750 unique compounds, including controls, which are described in a later section. These nominated compounds meet the following criteria:

- Structure of the compounds was in the public domain or could be released by the company.
- The purity of the compounds is either greater than 90% or is at least 80% with additional QC data.
- Lastly, the nominated compounds should not include any controlled substances, to facilitate sharing the compounds across international borders.

Genetic perturbation selection

ORF expression library

For cpg0016 we used a pre-existing lentiviral ORF expression library created at the Broad Institute for the gene overexpression experiments [22]. We tested 15,136 overexpression reagents encompassing 12,602 unique genes, including controls, which are described in a later section. Lentiviral packaging becomes less efficient for larger genes; this library does not cover the entire genome due to excluding larger genes.

CPJUMP1 data (cpg0000) also contains ORF perturbation data, where we overexpressed 176 genes. Additional details about these perturbations have been published [17].

CRISPR knockout library

We used the Horizon Discovery Human Edit-R™ synthetic crRNA - Druggable Genome in our CRISPR knockout experiments. This library targets 7975 unique genes, each gene is targeted by a pool of four predesigned synthetic crRNAs with high specificity and functional knockout.

In the CRISPR knockout experiments in CPJUMP1, we knocked down 160 genes. Additional details about these perturbations have been published [17].

Controls

We included several controls to identify and/or correct for different experimental artifacts. The controls can be broadly categorized as within-plate controls (those that are on the same plate as the treatments), and control plates (entire plates that are run alongside the treatment plates).

Within-plate controls

Within-plate controls include negative controls, positive control and untreated wells.

Negative control wells: These controls are used for detecting and correcting plate to plate variations. They can also be used as a baseline for identifying perturbations with a detectable morphological signal. The following are the within-plate negative controls.

- Compound plates – Dimethyl Sulfoxide (DMSO).
- ORF plates – ORFs of genes BFP, HcRed, lacZ and Luciferase, although it should be noted that no sensible negative control exists for protein overexpression, as each of these proteins is known to induce some changes to cell state.
- CRISPR plates – Non-targeting guides (CRISPR guides that do not target any gene) and DMSO.

Positive control wells: These controls are included in the plates to ensure that the experiment worked as expected.

- Compound plates – From the CPJUMP1 pilot experiment, we selected eight compounds with the most distinct signatures from each other and from DMSO. The identity of each compound is provided in Supplementary Table 1.
- ORF plates – We ran four of the eight compound positive controls on the ORF plates (Supplementary Table 1). Also, eGFP (enhanced Green Fluorescent Protein) is the ORF positive control.
- CRISPR plates – PLK1; knocking it down kills cells.

Untreated wells: Some wells were treated neither with treatments nor with controls; they contained only cells. These wells are unfortunate practical limitations, primarily useful for recognizing the plate orientation. They might also be used for normalizing samples, but they are non-ideal negative controls, given that they are not mock-treated with any reagents.

Control plates

Negative control plates: Some data acquisition partners ran negative control plates periodically (e.g., one per batch, or at the beginning and end of each batch). These plates can help disentangle staining and imaging artifacts, including batch effects and well-position effects. Examples include:

- An entire plate of untreated cells.
- An entire plate of cells treated with non-targeting guides (for CRISPR).
- An entire plate where all wells were treated with DMSO.

JUMP-Target-2-Compound plates: All partners ran one or more plates of JUMP-Target-2-Compound compounds in each batch to correct batch effects (unrelated plates from the same batch correlate strongly with each other compared to related plates in other batches).

JUMP-Target-1-Compound plates: To align the production data (cpg0016) with the CPJUMP1 experiment in the pilot data (cpg0000), we ran four JUMP-Target-1-Compound plates in one batch of cpg0016 from source_4 (Broad).

JUMP-Target-2-Compound plates with polybrene: To quantify the effects of adding polybrene, a viral transfection agent that is added to the ORF experiment, on the morphology of the cell, we ran two JUMP-Target-2-Compound plates with polybrene in one batch.

Plate layout design

Compound plates

The 384-well plate layouts (Supplementary Figure 7) were designed such that the controls are in the outer columns and the treatment wells are in the inner columns of the plate, to minimize well-position effects (also called plate-layout effects, whereby identical treatments in outer wells of a plate show different results than the same treatment positioned in inner ones) for the treatment samples, and to be compatible with partners' instrumentation and logistics. Four replicates of all eight compound positive controls are added to the outermost columns 1 and 24, while DMSO is added to the next-innermost columns, 2 and 23. The remaining wells contain one replicate of compound treatments. In the case of 1536-well plates (Supplementary Figure 8), each quadrant of the plate follows the same layout as a 384-well plate.

ORF plates

The ORF plates were pre-designed due to their existence in a pre-plated library [22]; each plate consists of negative control wells and untreated wells spread across the plate. Most plates contain 16 negative control wells, while some have as many as 28 wells (Supplementary Table 2). One replicate of four of the compound positive controls are added to wells O23, O24, P23 and P24. The remaining wells contain ORF treatments, with a single replicate of each per plate map and with five replicate plates produced per plate map.

CRISPR plates

Similar to the compound plates, the outer columns of the plate contain positive and negative controls. The outermost columns 1 and 24 contain four replicates of the eight compound positive controls. Columns 2 and 23 contain ten replicates each of wells with no-guides and non-targeting guides, eight replicates of DMSO and four replicates of the CRISPR positive control, PLK1.

JUMP-Target-1-Compound plates

The plates contain 306 compounds and DMSO; all but 14 compounds are in singlicates. There are 64 DMSO wells that are spread across the plate. The 14 compounds with two replicates are diverse positive controls that are different from the positive controls on the production plate. Additional details about these positive controls and the criteria met by the compounds of the plate have been described [17].

JUMP-Target-2-Compound plates

The contents of the plate are the same as that of JUMP-Target-1-Compound plates, but the layout is different (Supplementary Figure 9). JUMP-Target-2-Compound plates meet all the criteria met by JUMP-Target-1-Compound plates [1]. Both layouts are provided in our GitHub repository – https://github.com/jump-cellpainting/JUMP-Target).

Perturbation identifier

All compound, ORF and CRISPR perturbations are assigned a unique JUMP identifier. These identifiers start with the code *JCP2022_*.

Experimental conditions

We optimized several experimental parameters in the pilot experiments before embarking on data production; therefore, those experiments used a variety of parameters and did not all use the optimized final parameters. Experimental conditions of the pilot experiments and their optimization have been described [1,17]. The following are the experimental parameters used for generating the production data (cpg0016).

- Cell Painting dyes: We modified the concentrations of many dyes to improve their performance while simultaneously making the Cell Painting assay cheaper; the final choices are as described [1].
- Cell type: We compared A549 and U2OS cell lines in pilot experiments (cpg0000) [17] and settled on using U2OS because the performance of both cell lines were similar and there are previous datasets generated using U2OS, which allows comparison across experiments.
- Time point: We compared different time points for each perturbation modality (cpg0000) [17] and settled on the following time points: Compound – 48 hours, ORF – 48 hours, CRISPR – 96 hours.
- Reagent vendor: Given the similar performance of the Cell Painting dyes from both vendors, Thermo Fisher and PerkinElmer (cpg0001) [1], we performed our experiments with dyes from the latter, also known as PhenoVue™ Cell Painting Kit, 2.0 (part number PING22, PerkinElmer, Waltham, MA). PerkinElmer donated PhenoVue kits to the consortium.
- Microtiter plates: After comparing different plates for their ability to minimize evaporation in the outer wells (cpg0000) [1], we used PerkinElmer Cell Carrier Ultra for our data production. All sites used 384-well plates, except for source_1 and source_9 that used 1536-well plates.
- Microscope settings: Microscope settings are described in the sample preparation and image acquisition section below.

- Number of fields of view: We captured six (source_2 and source_10) or nine (source_3 through source_8, source_11, and source_15) fields of view in the 384-well plates and four fields of view in 1536-well plates (source_1 and source_9).
- Concentration of compounds: The treatment compounds were assayed at 10 uM at all sites, apart from source_7 where the compounds were assayed at 0.625 uM (the goal being to assay some of the compounds at a low concentration in addition to the higher concentration used for most of data production). The positive control compounds in compound, ORF and CRISPR plates were assayed at 5 uM. JUMP-Target-1-Compound and JUMP-Target-2-Compound plates were also assayed at 5 uM.
- ORF experiments: We followed previous protocols [23,24], with the following experimental parameters chosen: 5 replicates per virus plate, 1525 cells/384w, 30 uL media seeding volume/384w, 1 uL virus/384w and 4 ug/mL polybrene added 1 hour post-seeding, 30 minute spin at ~1000g, media change after 24 hrs removing polybrene and virus adding back 40 uL media, no selection with Blasticidin.
- Creation of a Cas9 cell line: We made U2OS-311 at the Broad Institute by transducing U2OS cells at a multiplicity of infection (MOI) <1 with lentivirus prepared from the vector pLX_311-Cas9 (Addgene plasmid 96924), which expresses blasticidin resistance from the SV40 promoter and Cas9 from the EF1α promoter, and selected with 16 µg/ml blasticidin for 14 d. Briefly, a 12-well plate at 1.5 × 10e6 cells per well in 1.25 mL media and 750 µL virus supplemented with 4 µg/ml polybrene were centrifuged for 2 h at 1,000*g,* then 2 mL media was added per well. 24 h after infection, cells were split out of the 12-well plate and 48 h after infection 16 µg/ml blasticidin was added and maintained. This is a slight modification of our published protocol [25].
- CRISPR experiments: 5 replicates per target gene were assessed, each well containing 4 different sequences targeting the same gene. The cell line used for these experiments was U2OS-311 (Broad). The cells were kept for a week in blasticidin before plating. Blasticidin was removed one passage before plating. Before plating, 125nL of each guide RNA at 10uM (Horizon Discovery) and 125nL of the tracrRNA at 10uM (Horizon Discovery) were dispensed in the 384 well plates using the ECHO (Beckman Coulter Life Science). 10uL of LipoRNAimax (Thermo Fisher) diluted in OptiMEM was added to each well using a MultiDrop Combi (Thermo Fisher) to achieve a final 0.05uL/well of LipoRNAimax. The plates were pulse centrifuged, then incubated for 30 minutes at room temperature. Then 700 cells per well were dispensed in 40uL using the MultiDrop Combi and incubated for 96h in an incubator (37C, 5% CO2) before staining and fixation. Additional chemical treatments were included in some control wells (DMSO, C1, C2, … C8) as well as full plates of cells treated with the Target-2 plate of compounds. These treatments were performed the same way as described in the chemical screening and were incubated 48h before staining and fixation on U2OS-311 cells with LipoRNAimax treatment only (no guide RNA, no tracrRNA).

Experiment design

The production data was generated at 12 sites or sources (Figure 1 and Supplementary Table 3). Compound data-producing partners are divided into two groups, wave 1 (7 partners) and wave 2 (3 partners). Additionally, there are two other data producing partners, one (source_4, the Broad) generated the 5-replicate ORF data and the other (source_13) generated a 5-replicate compound dataset and the 5-replicate CRISPR dataset. Each wave 1 partner shared their ~12,000 nominated compounds with four other wave 1 partners and received compounds from four other wave 1 partners. Thus, each partner screened roughly five times the number of compounds they nominated. Thus, each compound was typically screened in five replicates across five sites. Each wave 2 partner similarly shared their nominated compounds with the other two partners and received compounds from the other two partners. Each partner then screened three replicates of the compounds they nominated along with the compounds they received. Thus, each compound was screened in five replicates

across three sites. Nearly 6000 compounds were common between wave 1 and wave 2 partners, which will help align wave 1 and wave 2 compound production datasets. The plates were grouped into batches. The number of plates in each batch is different for different sources. In each batch, along with the treatment plates, most partners ran at least one untreated or DMSO plate and one JUMP-Target-2-Compound plate, so that the different batches can be aligned within a source and across sources.

Sample preparation and image acquisition

In the Cell Painting assay, eight cellular components are stained with six fluorescent dyes and imaged in five channels: mitochondria (MitoTracker; Mito), nucleus (Hoechst; DNA), nucleoli and cytoplasmic RNA (SYTO 14; RNA), endoplasmic reticulum (concanavalin A; ER), Golgi and plasma membrane (wheat germ agglutinin (WGA); AGP) and the actin cytoskeleton (phalloidin; AGP). We followed the optimized Cell Painting assay [1]; the differences between the v2 [26] and v3 protocols and any future updates are available at https://github.com/carpenterlab/2022_Cimini_NatureProtocols/wiki. Five types of microscope were used to acquire images at the 12 data production sites. These included the PerkinElmer Opera Phenix, ImageXpress Micro confocal, Yokogawa CV8000, Yokogawa CV7000, and PerkinElmer Operetta. Images were acquired in either widefield or confocal modes, and one or three brightfield planes were also acquired at some sites, apart from the fluorescent channels. The settings for each microscope and the filter configuration are provided in Supplementary Table 4 and Supplementary Table 6, respectively.

Image Processing

We used CellProfiler bioimage analysis software (versions 4.0.7, 4.1.3 or 4.2.1) to process the images. CellProfiler uses classical algorithms to measure cellular features. We corrected for variations in background illumination intensity and then segmented cells. We then measured several categories of features such as intensity, granularity, texture, and density for each cell in each fluorescent channel and each brightfield plane, if captured. Similarly, we measured features at the whole-image level. For more details, see http://broad.io/cellprofilermanual. Following the image analysis pipeline (see https://github.com/broadinstitute/imaging-platform-pipelines/tree/master/JUMP_production#production-pipelines for the pipelines), we measured up to 7648 features (including features measured at the per-cell level and those measured only at the per-image level) for images that included three brightfield planes which are each measured separately in the pipeline; for sources that collected fluorescence images only, the profiles contained 4762 features for each cell).

Image-based profiling

CellProfiler features were processed using pycytominer [27]. Single cell profiles were aggregated by calculating the mean profile for each sample on a plate. Each feature of the mean profile was then normalized by subtracting the median and dividing by the mean absolute deviation of the feature from either all the samples on the plate or only the negative control. Finally, the profiles are filtered to remove redundant features (features whose correlation with other features is higher than 0.9) and invariant features (features whose variance across all the samples is low).

Data availability

Images, single-cell profiles and all levels of well-level profiles (aggregated, annotated, normalized and feature selected) are available for cpg0000, cpg0001 and cpg0002. Only images, single cell profiles and aggregated well-level profiles from ten sources are available for cpg0016. For source_4, additional levels of well-level

profiles for cpg0016 are available at https://github.com/jump-cellpainting/jump-orf-data. The remaining data in cpg0016 will be released after we complete data validation. The details of which data is available and how to download it is provided at https://github.com/jump-cellpainting/datasets. Jupyter notebooks for generating the figures are available at https://github.com/jump-cellpainting/jump-data-production-paper/tree/6362150e822dbbdbf396b6dae21179cc802b0960/0.visualize-orf-data.

Data organization

The folder structure of the data repositories is provided in https://github.com/broadinstitute/cellpainting-gallery/blob/main/folder_structure.md.
Paper repository – https://github.com/jump-cellpainting/jump-data-production-paper

Tools and Software used

Figures were generated using python code written in a Jupyter notebook environment [28,29]. Plots were generated using matplotlib [30], seaborn [31] and Plotly [32] libraries along with Lucidchart and Google Draw. Two-dimensional representations of image-based profiles in Figures 2 and Supplementary Figures 1-6 were generated using UMAP [19] .

Acknowledgements

The authors appreciate the scientists and administrative personnel who contributed to the organization, scientific direction, and execution of the JUMP Cell Painting Consortium's project. In particular, we thank Justyna Różycka, Barbara Ściborowska, and Bartosz Szabucki, for work on the development of PhenAID JUMP-CP Data Explorer application, and Piotr Gaiński and Michał Koziarski for their efforts in preprocessing and analysis of the data presented in the application.

Funding

The authors gratefully acknowledge a grant from the Massachusetts Life Sciences Center Bits to Bytes Capital Call program for funding the data production and catalyzing this Consortium. We appreciate funding to support data analysis and interpretation from members of the JUMP Cell Painting Consortium (Amgen, AstraZeneca, Bayer AG, Biogen, Eisai, Janssen Pharmaceutica NV, Merck KGaA, Darmstadt, Germany, Pfizer, Servier, Takeda Development Center Americas, Inc. (TDCA)), from the National Institutes of Health (NIH MIRA R35 GM122547 to AEC), and from grant number 2020-225720 to BAC from the Chan Zuckerberg Initiative DAF, an advised fund of the Silicon Valley Community Foundation. We would like to acknowledge the Consortium's Supporting Partners for their in-kind contributions: Ardigen for their deep learning expertise and JUMP-CP Data Explorer web application (part of Ardigen's phenAID platform); Google/Verily for the compute support and configuration/optimization of Terra, which is co-developed by the Broad Institute of MIT and Harvard, Microsoft and Verily (its use is not described in this paper); Horizon Discovery, a PerkinElmer company, for the CRISPR-Cas9 library; Nomic bio for their protein profiling (not described in this paper); and PerkinElmer, for the PhenoVue™ Cell Painting JUMP kit. We also are grateful for the Amazon Web Services Registry of Open Data for hosting the public dataset. The authors also gratefully acknowledge the use of the PerkinElmer Opera Phenix® High-Content/High-Throughput imaging system at the Broad Institute, funded by the S10 Grant NIH OD-026839.

References

1. Cimini BA, Chandrasekaran SN, Kost-Alimova M. Optimizing the Cell Painting assay for image-based profiling. bioRxiv [Internet]. biorxiv.org; 2022; Available from: https://www.biorxiv.org/content/10.1101/2022.07.13.499171.abstract

2. Chandrasekaran SN, Ceulemans H, Boyd JD, Carpenter AE. Image-based profiling for drug discovery: due for a machine-learning upgrade? Nat Rev Drug Discov. Nature Publishing Group; 2020;1–15.

3. Schiff L, Migliori B, Chen Y, Carter D, Bonilla C, Hall J, Fan M, Tam E, Ahadi S, Fischbacher B, Geraschenko A, Hunter CJ, Venugopalan S, DesMarteau S, Narayanaswamy A, Jacob S, Armstrong Z, Ferrarotto P, Williams B, Buckley-Herd G, Hazard J, Goldberg J, Coram M, Otto R, Baltz EA, Andres-Martin L, Pritchard O, Duren-Lubanski A, Daigavane A, Reggio K, NYSCF Global Stem Cell Array® Team, Nelson PC, Frumkin M, Solomon SL, Bauer L, Aiyar RS, Schwarzbach E, Noggle SA, Monsma FJ Jr, Paull D, Berndl M, Yang SJ, Johannesson B. Integrating deep learning and unbiased automated high-content screening to identify complex disease signatures in human fibroblasts. Nat Commun. 2022 Mar 25;13(1):1590. PMCID: PMC8956598

4. Cuccarese MF, Earnshaw BA, Heiser K, Fogelson B, Davis CT, McLean PF, Gordon HB, Skelly KR, Weathersby FL, Rodic V, Quigley IK, Pastuzyn ED, Mendivil BM, Lazar NH, Brooks CA, Carpenter J, Probst BL, Jacobson P, Glazier SW, Ford J, Jensen JD, Campbell ND, Statnick MA, Low AS, Thomas KR, Carpenter AE, Hegde SS, Alfa RW, Victors ML, Haque IS, Chong YT, Gibson CC. Functional immune mapping with deep-learning enabled phenomics applied to immunomodulatory and COVID-19 drug discovery [Internet]. bioRxiv. 2020 [cited 2021 Dec 1]. p. 2020.08.02.233064. Available from: https://www.biorxiv.org/content/10.1101/2020.08.02.233064v2

5. Moshkov N, Becker T, Yang K, Horvath P, Dancik V. Predicting compound activity from phenotypic profiles and chemical structures. bioRxiv [Internet]. biorxiv.org; 2022; Available from: https://www.biorxiv.org/content/10.1101/2020.12.15.422887.abstract

6. Hofmarcher M, Rumetshofer E, Clevert DA, Hochreiter S, Klambauer G. Accurate Prediction of Biological Assays with High-Throughput Microscopy Images and Convolutional Networks. J Chem Inf Model. 2019 Mar 25;59(3):1163–1171. PMID: 30840449

7. Trapotsi MA, Mouchet E, Williams G, Monteverde T, Juhani K, Turkki R, Miljković F, Martinsson A, Mervin L, Müllers E, Barrett I, Engkvist O, Bender A, Moreau K. Cell morphological profiling enables high-throughput screening for PROteolysis TArgeting Chimera (PROTAC) phenotypic signature [Internet]. bioRxiv. 2022 [cited 2022 Mar 5]. p. 2022.01.17.476610. Available from: https://www.biorxiv.org/content/10.1101/2022.01.17.476610v1.full

8. Nyffeler J, Willis C, Lougee R, Richard A, Paul-Friedman K, Harrill JA. Bioactivity screening of environmental chemicals using imaging-based high-throughput phenotypic profiling. Toxicol Appl Pharmacol. 2020 Jan 15;389:114876. PMID: 31899216

9. Way GP, Natoli T, Adeboye A, Litichevskiy L, Yang A, Lu X, Caicedo JC, Cimini BA, Karhohs K, Logan DJ, Rohban MH, Kost-Alimova M, Hartland K, Bornholdt M, Chandrasekaran SN, Haghighi M, Weisbart E, Singh S, Subramanian A, Carpenter AE. Morphology and gene expression profiling provide complementary information for mapping cell state [Internet]. bioRxiv. 2022 [cited 2022 Mar 17]. p. 2021.10.21.465335. Available from: https://www.biorxiv.org/content/10.1101/2021.10.21.465335

10. Rohban MH, Singh S, Wu X, Berthet JB, Bray MA, Shrestha Y, Varelas X, Boehm JS, Carpenter AE. Systematic morphological profiling of human gene and allele function via Cell Painting. Elife [Internet]. 2017 Mar 18;6. Available from: http://dx.doi.org/10.7554/eLife.24060 PMCID: PMC5386591

11. Moshkov N, Bornholdt M, Benoit S, Smith M, McQuin C, Goodman A, Senft RA, Han Y, Babadi M, Horvath P, Cimini BA, Carpenter AE, Singh S, Caicedo JC. Learning representations for image-based profiling of

perturbations [Internet]. bioRxiv. 2022 [cited 2022 Oct 14]. p. 2022.08.12.503783. Available from: https://www.biorxiv.org/content/10.1101/2022.08.12.503783

12. Cross-Zamirski JO, Williams G, Mouchet E, Schönlieb CB, Turkki R, Wang Y. Self-Supervised Learning of Phenotypic Representations from Cell Images with Weak Labels [Internet]. arXiv [cs.CV]. 2022. Available from: http://arxiv.org/abs/2209.07819

13. Williams E, Moore J, Li SW, Rustici G, Tarkowska A, Chessel A, Leo S, Antal B, Ferguson RK, Sarkans U, Brazma A, Salas REC, Swedlow JR. The Image Data Resource: A Bioimage Data Integration and Publication Platform. Nat Methods. 2017 Aug;14(8):775–781. PMCID: PMC5536224

14. Ljosa V, Sokolnicki KL, Carpenter AE. Annotated high-throughput microscopy image sets for validation. Nat Methods. 2012 Jun 28;9(7):637. PMCID: PMC3627348

15. Bray MA, Gustafsdottir SM, Rohban MH, Singh S, Ljosa V, Sokolnicki KL, Bittker JA, Bodycombe NE, Dancík V, Hasaka TP, Hon CS, Kemp MM, Li K, Walpita D, Wawer MJ, Golub TR, Schreiber SL, Clemons PA, Shamji AF, Carpenter AE. A dataset of images and morphological profiles of 30 000 small-molecule treatments using the Cell Painting assay. Gigascience. 2017 Dec 1;6(12):1–5. PMCID: PMC5721342

16. Caicedo JC, Arevalo J, Piccioni F, Bray MA, Hartland CL, Wu X, Brooks AN, Berger AH, Boehm JS, Carpenter AE, Singh S. Cell Painting predicts impact of lung cancer variants. Mol Biol Cell. 2022 May 15;33(6):ar49. PMID: 35353015

17. Chandrasekaran SN, Cimini BA, Goodale A, Miller L, Kost-Alimova M, Jamali N, Doench J, Fritchman B, Skepner A, Melanson M, Arevalo J, Caicedo JC, Kuhn D, Hernandez D, Berstler J, Shafqat-Abbasi H, Root D, Swalley S, Singh S, Carpenter AE. Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations [Internet]. bioRxiv. 2022 [cited 2022 Nov 18]. p. 2022.01.05.475090. Available from: https://www.biorxiv.org/content/10.1101/2022.01.05.475090v1

18. Jamali N, Tromans-Coia C, Abbasi HS, Giuliano KA, Hagimoto M, Jan K, Kaneko E, Letzsch S, Schreiner A, Sexton JZ, Suzuki M, Joseph Trask O, Yamaguchi M, Yanagawa F, Yang M, Carpenter AE, Cimini BA. Assessing the performance of the Cell Painting assay across different imaging systems [Internet]. bioRxiv. 2023 [cited 2023 Feb 21]. p. 2023.02.15.528711. Available from: https://www.biorxiv.org/content/10.1101/2023.02.15.528711v1

19. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction [Internet]. arXiv [stat.ML]. 2018. Available from: http://arxiv.org/abs/1802.03426

20. Tan M, Le QV. EfficientNetV2: Smaller Models and Faster Training [Internet]. arXiv [cs.CV]. 2021. Available from: http://arxiv.org/abs/2104.00298

21. Luecken MD, Büttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, Strobl DC, Zappia L, Dugas M, Colomé-Tatché M, Theis FJ. Benchmarking atlas-level data integration in single-cell genomics [Internet]. Nat. Methods. 2020 [cited 2021 Dec 6]. p. 2020.05.22.111161. Available from: http://dx.doi.org/10.1038/s41592-021-01336-8 PMCID: PMC8748196

22. Yang X, Boehm JS, Yang X, Salehi-Ashtiani K, Hao T, Shen Y, Lubonja R, Thomas SR, Alkan O, Bhimdi T, Green TM, Johannessen CM, Silver SJ, Nguyen C, Murray RR, Hieronymus H, Balcha D, Fan C, Lin C, Ghamsari L, Vidal M, Hahn WC, Hill DE, Root DE. A public genome-scale lentiviral expression library of human ORFs. Nat Methods. 2011 Jun 26;8(8):659–661. PMCID: PMC3234135

23. Johannessen CM, Boehm JS, Kim SY, Thomas SR, Wardwell L, Johnson LA, Emery CM, Stransky N, Cogdill AP, Barretina J, Caponigro G, Hieronymus H, Murray RR, Salehi-Ashtiani K, Hill DE, Vidal M, Zhao JJ, Yang X, Alkan O, Kim S, Harris JL, Wilson CJ, Myer VE, Finan PM, Root DE, Roberts TM, Golub T, Flaherty KT, Dummer R, Weber BL, Sellers WR, Schlegel R, Wargo JA, Hahn WC, Garraway LA. COT drives resistance to RAF inhibition through MAP kinase pathway reactivation. Nature. 2010 Dec 16;468(7326):968–972. PMCID: PMC3058384

24. Berger AH, Brooks AN, Wu X, Shrestha Y, Chouinard C, Piccioni F, Bagul M, Kamburov A, Imielinski M, Hogstrom L, Zhu C, Yang X, Pantel S, Sakai R, Watson J, Kaplan N, Campbell JD, Singh S, Root DE, Narayan R, Natoli T, Lahr DL, Tirosh I, Tamayo P, Getz G, Wong B, Doench J, Subramanian A, Golub TR, Meyerson M, Boehm JS. High-throughput Phenotyping of Lung Cancer Somatic Mutations. Cancer Cell. 2016 Aug 8;30(2):214–228. PMCID: PMC5003022

25. Doench JG, Hartenian E, Graham DB, Tothova Z, Hegde M, Smith I, Sullender M, Ebert BL, Xavier RJ, Root DE. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. Nat Biotechnol. 2014 Dec;32(12):1262–1267. PMCID: PMC4262738

26. Bray MA, Singh S, Han H, Davis CT, Borgeson B, Hartland C, Kost-Alimova M, Gustafsdottir SM, Gibson CC, Carpenter AE. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. Nat Protoc. Nature Research; 2016 Sep;11(9):1757–1774. PMCID: PMC5223290

27. Way G, Chandrasekaran SN, Bornholdt M, Fleming S, Tsang H, Adeboye A, Cimini B, Weisbart E, Ryder P, Stirling D, Jamali N, Carpenter A, Singh S. Pycytominer: Data processing functions for profiling perturbations [Internet]. Available from: https://github.com/cytomining/pycytominer

28. Loizides F, Schmidt B. Positioning and Power in Academic Publishing: Players, Agents and Agendas: Proceedings of the 20th International Conference on Electronic Publishing. IOS Press; 2016.

29. Van Rossum, Drake. The python language reference. Python software foundation [Internet]. Available from: https://dev.rbcafe.com/python/python-3.5.1-pdf/reference.pdf

30. Hunter. Matplotlib: A 2D Graphics Environment. 2007 May 1;9:90–95.

31. Waskom M. seaborn: statistical data visualization. J Open Source Softw. The Open Journal; 2021 Apr 6;6(60):3021.

32. Inc PT. Collaborative data science. Montreal: Plotly Technologies Inc Montréal [Internet]. 2015; Available from: https://plot.ly