RxRx1:
An image set for cellular morphological variation across many experimental batches

James Taylor*, **Berton Earnshaw***, **Ben Mabey***, **Mason Victors*** & **Jason Yosinski***,†
*Recursion Pharmaceuticals, †Uber AI Labs
{james.taylor,berton.earnshaw,ben.mabey,mason.victors}@recursionpharma.com, yosinski@uber.com

Abstract

High-throughput screening techniques are commonly used in many fields of biology. However, it is well known that non-biological artifacts arising from variability in the technical execution of different experimental batches confound high-throughput screens measurements. These *batch effects* obscure biological conclusions, and it is therefore necessary to account for them. While a number of techniques have been proposed, to our knowledge there is not a publicly available biological dataset designed specifically for the systematic study of batch effect correction. To this end we announce the release of RxRx1, a set of 125,514 high-resolution fluorescence microscopy images of human cells under 1,108 genetic perturbations in 51 experimental batches across four cell types. Visual inspection of the images by batch makes it clear that the set indeed demonstrates significant batch effects. In this paper we describe the image set in detail. We also propose a classification task designed to study batch effect correction on these images, and provide some baseline results for the task. Our goal in releasing this image set is to encourage researchers across various disciplines to develop effective methods for removing batch effects that generalize well to unseen experimental batches and to share these methods with the scientific community.

1 Introduction

High-throughput screening techniques are in common use in many biological fields, including genetics (Echeverri & Perrimon, 2006; Zhou et al., 2014) and drug discovery (Broach et al., 1996; Macarron et al., 2011; Swinney & Anthony, 2011; Boutros et al., 2015). Such techniques are capable of generating large amounts of data that, when coupled with modern machine learning methods, could help in answering fundamental questions in biology. These techniques may also help ameliorate the problem of the exponential rise in the cost of developing an approved drug, which is now estimated to be well over $2 billion (Scannell et al., 2012; DiMasi et al., 2016). However, creating such large volumes of biological data necessarily requires the data to be generated in experimental batches, or groups of experiments executed at similar times under similar conditions. Even when experiments are carefully designed to control for technical variables such as temperature, humidity, and reagent concentration, the measurements taken from these screens are confounded by non-biological artifacts that arise from variability in the technical execution of each batch. These *batch effects* create factors of variation within the data that are irrelevant to the biological variables under study, but are unfortunately often correlated with them. It is therefore necessary to correct for batch effects before drawing any biological conclusions from measurements taken from high-throughput screens (Leek et al., 2010; Parker & Leek, 2012; Soneson et al., 2014; Nygaard et al., 2016).

Many computational methods have been designed for dealing with batch effects (Leek et al., 2010; Chen et al., 2011; Lazar et al., 2012; Parker & Leek, 2012; Leek et al., 2012; Goh et al., 2017; Shaham et al., 2017), yet to our knowledge there are no publicly-available biological datasets that were systematically created to study them. Here we announce the public release of such a dataset, which we call RxRx1. The dataset consists of images of human cells under more than 1,100 different genetic perturbations across 51 experimental batches and four cell types. We also propose a machine learning task that gauges the effectiveness of the batch effect correction method — correctly classify

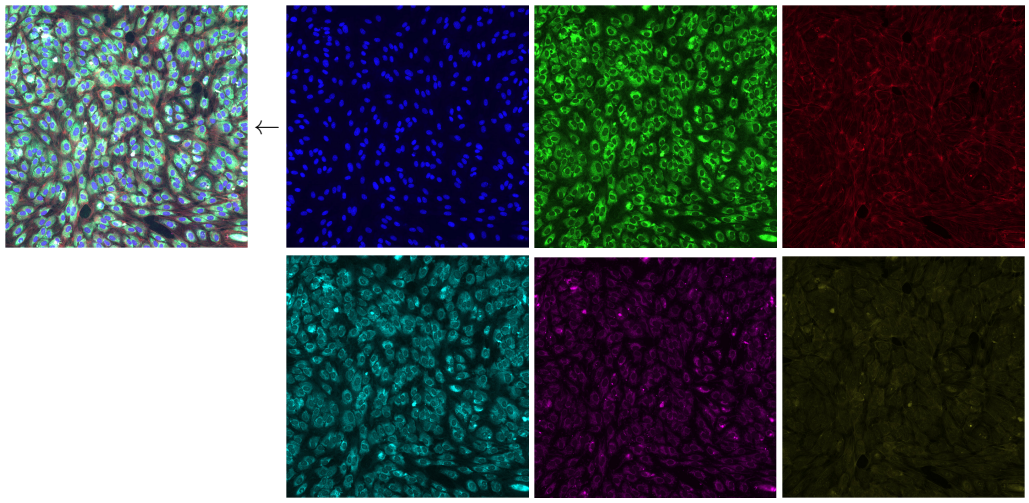

Figure 1: 6-channel faux-colored composite image of HUVEC cells (left) and individual channels (rest): nuclei (blue), endoplasmic reticuli (green), actin (red), nucleoli and cytoplasmic RNA (cyan), mitochondria (magenta), and Golgi (yellow). The similarity in content between some channels is due in part to the spectral overlap between the fluorescent stains used in those channels.

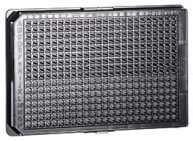

Figure 2: A 384-well plate. Experiments used to generate the images in this dataset were run in the wells of such plates. Photo courtesy of Greiner Bio One International GmbH.

the genetic perturbation present in each image in a held-out set of batches. In order for the classifier to generalize to unseen batches, it must learn to separate biological and technical factors in test images and make predictions only on the biological factors.

This dataset and task will be of interest to the rapidly growing community of researchers applying machine learning methods to complex biological data sets, especially those working with high-content phenotypic screens (Angermueller et al., 2016; Kraus et al., 2016; Caicedo et al., 2017; Kraus et al., 2017; Ando et al., 2017; Chen et al., 2018). The specific task of removing batch effects is relevant to the broader life sciences community and can provide insights that enable researchers to develop improved methods for working with other biological datasets. In addition, we hope the dataset is of interest to the larger community of machine learning researchers working in computer vision, especially those in the areas of domain adaptation, transfer learning, and k-shot learning.

2 Description of the Dataset

The image set was produced by Recursion Pharmaceuticals in its automated high-throughput screening laboratory. It is comprised of fluorescence microscopy images of human cells of four different types — HUVEC, RPE, HepG2, and U2OS — which were acquired using a 6-channel variation of the *Cell Painting* imaging protocol (Bray et al., 2016). In Figure 1, we show an example image.

The six channels of an image illuminate the different parts of the cell population in the field of view: nuclei, endoplasmic reticuli, actin, nucleoli and cytoplasmic RNA, mitochondria, and Golgi. The images themselves are the result of running 51 different instances of the same type of experiment. Each experiment instance is comprised of four 384-well plates (see Fig. 2), used to isolate populations of cells into wells. The wells are laid out on each plate in a 16×24 grid, but only the

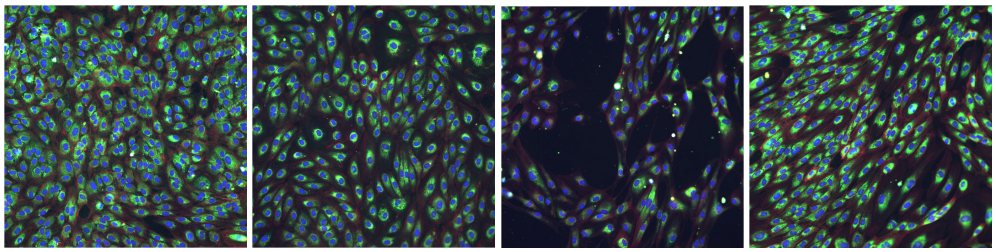

Figure 3: Images of four different siRNA phenotypes in HUVEC (same experiment and plate).

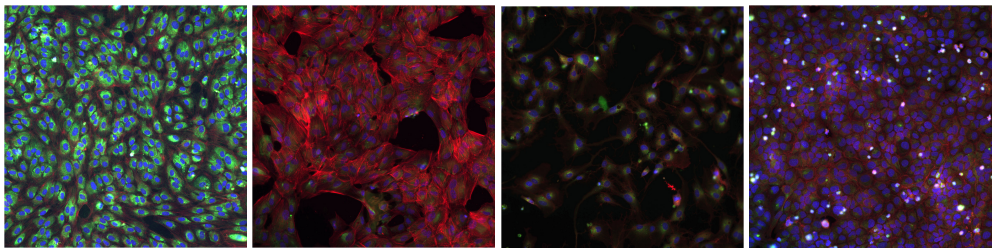

Figure 4: Images of the same siRNA in four cell types: HUVEC, RPE, HepG2, U2OS.

wells in the inner 14×22 grid are used since the outer wells are most susceptible to environmental factors. Of these 308 usable wells, one remains untreated to provide a negative control. The rest of the 307 wells receive exactly one small interfering ribonucleic acid, or siRNA, at a fixed concentration. Each siRNA is designed to knockdown a single target gene via the RNA interference pathway, reducing the expression of the gene and its associated protein (Tuschl, 2001). However, siRNAs are known to have significant but consistent off-target effects via the microRNA pathway, creating partial knockdown of many other genes as well. The overall effect of siRNA transfection is to perturb the morphology, count, and distribution of cells in each well, creating a distinct *phenotype* associated with each siRNA. The phenotype is sometimes visually recognizable from the images, but often the specific difference in cell morphology is subtle and hard to detect to the human eye (see Fig. 3).

In each experiment, the same 30 siRNA appear on every plate as a control set for the plate. These control siRNA target different genes and produce a variety of phenotypic effects that, taken in combination with the single untreated well, provide a set of useful reference wells for each plate. The 1,108 remaining wells of each experiment (277 wells×4 plates) receive 1,108 different siRNA. These non-control siRNA target different genes than each other and the genes of the control siRNA. Notice that while the control siRNA appear on each plate, each non-control siRNA appears at most once in each experiment. We say at most once because, although rare, it happens that either an siRNA is not transferred into its well, resulting in an additional untreated well on the plate, or an operational error is detected by quality control procedures and renders the well unsuitable for inclusion in the dataset.

When the images were originally acquired from the microscope, they were of spatial resolution 2048×2048, but in order to make the dataset more manageable, they were downsampled to 1024×1024 and cropped to the center 512×512 field of view. The image set contains two non-overlapping 512×512 fields of view per well. Therefore, there could be as many as 125,664 images (= 51 experiments×4 plates/experiment×4 wells/plate×2 images/well), but, because of operational errors, a number of images were removed, resulting in 125,514 actual images in the dataset.

As was mentioned, the entire dataset consists of 51 experiments: 24 in HUVEC, 11 in RPE, 11 in HepG2, and 5 in U2OS. Figure 4 shows the phenotype of a single siRNA in the four different cell types. Each of the 51 experiments was run in a different batch, resulting in images that exhibit technical effects (e.g. differences in temperature, humidity, siRNA concentration) that are common to the batch but distinct from other batches (see Fig. 5). It is this feature of the dataset that makes it particularly suited for studying batch effects and methods for correcting them.

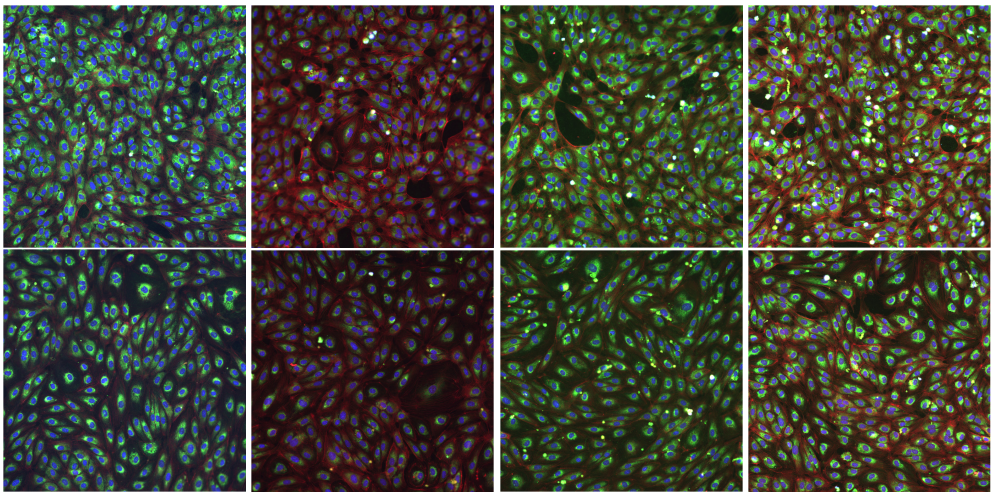

Figure 5: Images of two different siRNA (rows) in HUVEC cells across four experimental batches (columns). Notice the visual similarity of images from the same batch.

Table 1: Average test accuracies for all cell types and per cell type.

Split	All	HUVEC	RPE	HepG2	U2OS
Batch	44.1% ± 5.7	56.7% ± 8.4	31.2% ± 2.7	30.8% ± 4.0	2.4% ± 1.0
Random	52.8% ± 0.0	68.9% ± 0.3	40.0% ± 0.9	39.7% ± 0.2	22.0% ± 0.4

The image set is accompanied by metadata providing the following information about each image: cell type, experiment, plate, well location, and treatment class (1,138 siRNA classes plus one untreated class).

3 Proposed task for studying batch effect correction

While the dataset is useful for many types of studies (e.g. domain adaptation, k-shot learning), we propose the following task for studying batch effect correction: correctly classify the images of non-control siRNA in a hold-out set of batches. In order for a classifier to generalize well to unseen batches, it must learn to separate biological factors associated with siRNA perturbation from technical factors associated with batch effects in the training batches, and use the biological factors for classification. We illustrate this point with the following experiment. We subset the data to 42 experiments (20 HUVEC, 9 RPE, 9 HepG2, 4 U2OS), and randomly chose 33 experiments for training (16 HUVEC, 7 RPE, 7 HepG2, 3 U2OS) and 9 for testing (4 HUVEC, 2 RPE, 2 HepG2, 1 U2OS) and trained a standard ResNet50 (He et al., 2016) on just the 1,108 non-control siRNA in the training set. The image intensities were standardized by the means and standard deviations of the control intensities per channel per plate. While training accuracies all reached near 100%, the average test accuracy over three such batch splits of the data was 44.1%. To assess the extent to which batch effects affected these results, we randomly split the 42 experiments again into training and test sets of similar sizes, but without accounting for experiment so that samples from each experiment appear in both the training and test sets. The average test accuracy in this case was 52.8%, or 20% higher than when we split by batch. Table 3 summarizes these results, including test accuracies of models trained on individual cell lines. Using this dataset, we hope researchers will design novel methods for correcting batch effects and benchmark themselves against this classification task. To promote this research, we are running a competition for this dataset and task later this year.

REFERENCES

D Michael Ando, Cory McLean, and Marc Berndl. Improving phenotypic measurements in high-content imaging screens. *bioRxiv*, pp. 161422, 2017.

Christof Angermueller, Tanel Pärnamaa, Leopold Parts, and Oliver Stegle. Deep learning for computational biology. *Molecular systems biology*, 12(7):878, 2016.

Michael Boutros, Florian Heigwer, and Christina Laufer. Microscopy-based high-content screening. *Cell*, 163(6):1314–1325, 2015.

Mark-Anthony Bray, Shantanu Singh, Han Han, Chadwick T Davis, Blake Borgeson, Cathy Hartland, Maria Kost-Alimova, Sigrun M Gustafsdottir, Christopher C Gibson, and Anne E Carpenter. Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature protocols*, 11(9):1757, 2016.

James R Broach, Jeremy Thorner, et al. High-throughput screening for drug discovery. *Nature*, 384 (6604):14–16, 1996.

Juan C Caicedo, Sam Cooper, Florian Heigwer, Scott Warchal, Peng Qiu, Csaba Molnar, Aliaksei S Vasilevich, Joseph D Barry, Harmanjit Singh Bansal, Oren Kraus, et al. Data-analysis strategies for image-based cell profiling. *Nature methods*, 14(9):849, 2017.

Chao Chen, Kay Grennan, Judith Badner, Dandan Zhang, Elliot Gershon, Li Jin, and Chunyu Liu. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PloS one*, 6(2):e17238, 2011.

Hongming Chen, Ola Engkvist, Yinhai Wang, Marcus Olivecrona, and Thomas Blaschke. The rise of deep learning in drug discovery. *Drug discovery today*, 23(6):1241–1250, 2018.

Joseph A DiMasi, Henry G Grabowski, and Ronald W Hansen. Innovation in the pharmaceutical industry: new estimates of r&d costs. *Journal of health economics*, 47:20–33, 2016.

Christophe J Echeverri and Norbert Perrimon. High-throughput rnai screening in cultured cells: a user's guide. *Nature Reviews Genetics*, 7(5):373, 2006.

Wilson Wen Bin Goh, Wei Wang, and Limsoon Wong. Why batch effects matter in omics data, and how to avoid them. *Trends in biotechnology*, 35(6):498–507, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Oren Z Kraus, Jimmy Lei Ba, and Brendan J Frey. Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics*, 32(12):i52–i59, 2016.

Oren Z Kraus, Ben T Grys, Jimmy Ba, Yolanda Chong, Brendan J Frey, Charles Boone, and Brenda J Andrews. Automated analysis of high-content microscopy data with deep learning. *Molecular systems biology*, 13(4):924, 2017.

Cosmin Lazar, Stijn Meganck, Jonatan Taminau, David Steenhoff, Alain Coletta, Colin Molter, David Y Weiss-Solís, Robin Duque, Hugues Bersini, and Ann Nowé. Batch effect removal methods for microarray gene expression data integration: a survey. *Briefings in bioinformatics*, 14(4): 469–490, 2012.

Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733, 2010.

Jeffrey T Leek, W Evan Johnson, Hilary S Parker, Andrew E Jaffe, and John D Storey. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6):882–883, 2012.

Ricardo Macarron, Martyn N Banks, Dejan Bojanic, David J Burns, Dragan A Cirovic, Tina Garyantes, Darren VS Green, Robert P Hertzberg, William P Janzen, Jeff W Paslay, et al. Impact of high-throughput screening in biomedical research. *Nature reviews Drug discovery*, 10(3):188, 2011.

Vegard Nygaard, Einar Andreas Rødland, and Eivind Hovig. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics*, 17(1):29–39, 2016.

Hilary S Parker and Jeffrey T Leek. The practical effect of batch on genomic prediction. *Statistical applications in genetics and molecular biology*, 11(3), 2012.

Jack W Scannell, Alex Blanckley, Helen Boldon, and Brian Warrington. Diagnosing the decline in pharmaceutical r&d efficiency. *Nature reviews Drug discovery*, 11(3):191, 2012.

Uri Shaham, Kelly P Stanton, Jun Zhao, Huamin Li, Khadir Raddassi, Ruth Montgomery, and Yuval Kluger. Removal of batch effects using distribution-matching residual networks. *Bioinformatics*, 33(16):2539–2546, 2017.

Charlotte Soneson, Sarah Gerster, and Mauro Delorenzi. Batch effect confounding leads to strong bias in performance estimates obtained by cross-validation. *PloS one*, 9(6):e100335, 2014.

David C Swinney and Jason Anthony. How were new medicines discovered? *Nature reviews Drug discovery*, 10(7):507, 2011.

Thomas Tuschl. Rna interference and small interfering rnas. *Chembiochem*, 2(4):239–245, 2001.

Yuexin Zhou, Shiyou Zhu, Changzu Cai, Pengfei Yuan, Chunmei Li, Yanyi Huang, and Wensheng Wei. High-throughput screening of a crispr/cas9 library for functional genomics in human cells. *Nature*, 509(7501):487, 2014.