



Software Manual – v1.2.1

High-throughput image analysis software using *SDCubes* for single-cell
data storage and adaptive experimental design

Bjorn L Millard¹, Mario Niepel¹, Michael P Menden^{1,3}, Jeremy L Muhlich¹,
and Peter K Sorger^{1,2}

January 11, 2011

¹Center for Cell Decision Processes, Department of Systems Biology, Harvard Medical School, Boston,
MA 02115, USA, ²Department of Biological Engineering, Massachusetts Institute of Technology,
Cambridge, MA 02139, USA ³Current address: University of Applied Sciences Weihenstephan-Triesdorf



This manual © Copyright 2011 - Bjorn Millard

This manual licensed under the Creative Commons Attribution-NoDerivs 3.0 United States License

Email: imagerail-users@googlegroups.com
Web: <http://www.semanticbiology.com/software/imagerail>

License

ImageRail: Software for high-throughput microscopy image analysis

Copyright (C) 2011 Bjorn Millard <bjornmillard@gmail.com>

This program is free software: you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation, either version 3 of the License, or (at your option) any later version.

This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.

You should have received a copy of the GNU General Public License along with this program. If not, see <<http://www.gnu.org/licenses/>>.

Contents

- 1 Introduction
- 2 Installation and Quick Start
- 3 *ImageRail* Projects
 - 3.1 Creating New Project
 - 3.2 Opening Existing Project
- 4 Images
 - 4.1 Image Formats and Conventions
 - 4.2 Loading Images
 - 4.3 Image Viewing
 - 4.3.1 Field Viewer
 - 4.3.2 Montage Viewer
- 5. Image Processing
 - 5.1 Well Means
 - 5.2 Single-Cell Segmentation
 - 5.2.1 Load cells into RAM after segmentation
 - 5.2.2 Coordinates for HDF5 Storage
 - 5.3 Parameter Selection
- 6. Data Viewing
 - 6.1 Plate Graph
 - 6.2 Line Plot
 - 6.3 Dot Plot
 - 6.4 Histogram Plot
- 7. Data Export
 - 7.1 CSV – Well Means
 - 7.2 MIDAS
 - 7.2.1 Well Means
 - 7.2.2 Single-Cells
- 8. Experimental Design Metadata
- 9 Acknowledgments

1 Introduction

This short document is a manual for *ImageRail* software first described by Millard et al. [1]. Additional resources and periodic updates can be found at: <http://www.semanticbiology.com/software>. This website also offers further details on SDCubes, the storage format used by ImageRail, including a separate manual and software library.

The goal of *ImageRail* software is to provide an open-source, stand-alone program built around the *SDCube* data structure that 1) facilitates high-throughput image analysis, including segmentation and feature computation, 2) provides basic data exploration tools such as line plots, scatter plots, histograms and heat maps, and 3) encodes and manipulates experimental design metadata and links it to extracted data from the source images.

ImageRail data and metadata are stored in *semantically-typed data hypercubes (SDCubes)*, a data structure built from the HDF5 and XML file formats. *SDCubes* were developed to efficiently store large amounts of experimental biology data (in this case, single-cell data

acquired via high-throughput microscopy; HTM) and to link it to digitized experimental metadata describing how each sample was perturbed experimentally and then measured.

Although *ImageRail* data is stored primarily in *SDCubes*, user-defined subsets of the data can be exported in CSV format for analysis in other software programs such as *Excel*, *MatLab* and *Spotfire*.

The default *ImageRail* segmentation algorithm implements a simple watershed-style method. Currently, alternative algorithms can be substituted by performing some *Java* programming. We are currently creating new programming APIs that will create a direct interface with open-source image analysis programs *ImageJ*.

The full benefits of *SDCubes* will become apparent when more programs can read and

manipulate HDF5-encoded data and the XML-encoded experimental design data to analyze and navigate efficiently through the digitized single-cell data. Currently, *ImageRail* can create, modify and read *SDCubes*, but more advanced functional analysis and plotting of *SDCubes* will be handled by a sister program, *DataRail v2.0*, which is currently under development [2].

2 Installation and Quick Start

ImageRail downloads can be found at: <http://www.semanticbiology.com/software/imagerail>

Windows:

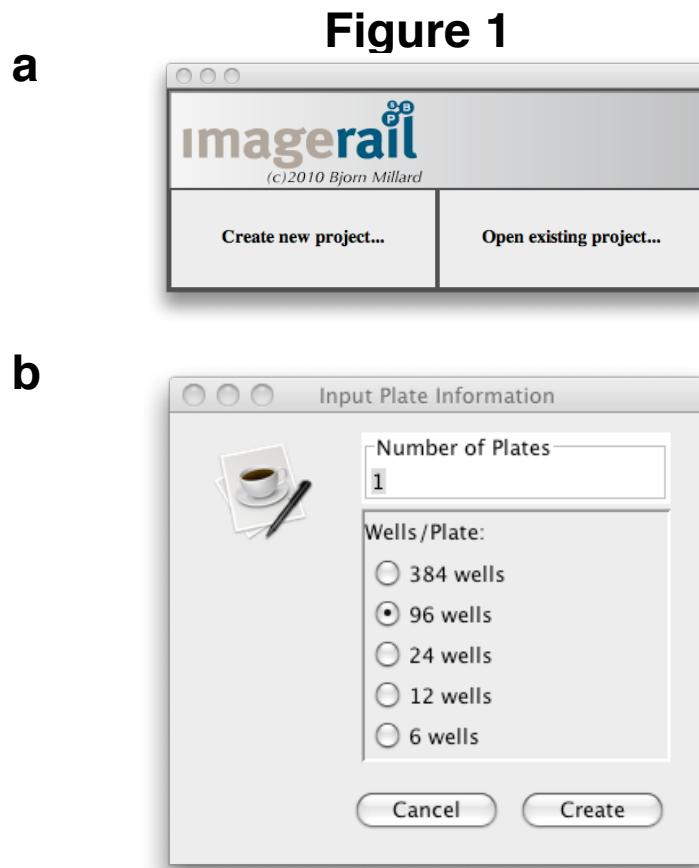
- 1) Download the Windows installer .exe file
- 2) Double click the installer file to initiate installation
- 3) Optional: To make a Desktop shortcut, drag the ImageRail icon from the Start menu Programs folder to your desktop while holding the Ctrl key.
- 4) ImageRail may be launched from the Start menu's Programs folder, or your desktop shortcut if you chose to create one.

Mac OS X (OS X 10.6 - Snow Leopard or greater is required)

- 1) Download the Mac OS X .dmg
- 2) Double-click the .dmg to mount it and display its contents
- 3) Drag the ImageRail application to your Applications folder
- 4) Optional: Drag the instruction manual PDF file to your Documents folder or wherever you would like to keep it.
- 5) To launch ImageRail, open your Applications folder and double-click the ImageRail application.

3 The *ImageRail* Project

After the *ImageRail* splash screen goes away, a dialog box should appear (Figure 1a). This box will let you create a new *ImageRail* project or re-open an existing project.



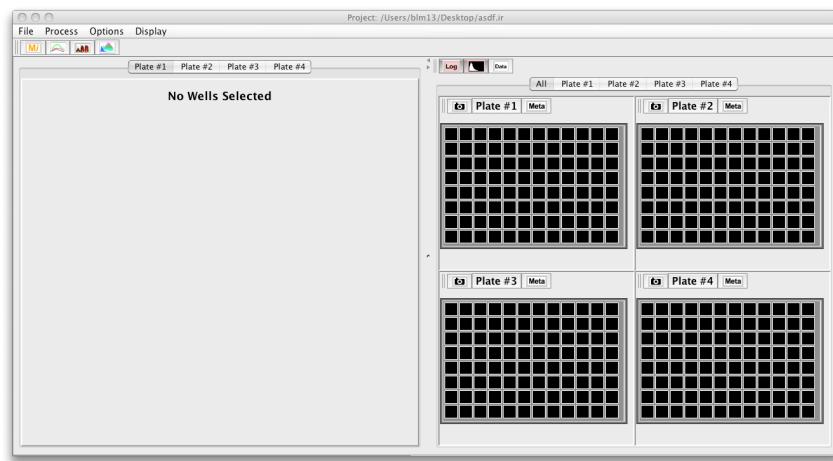
3.1 Creating New Projects

To create a new project, you must specify how many plates and how many wells per plate you want to initialize using the appropriate dialog box (Figure 1b).

3.2 Opening Existing Projects

If you want to re-open an existing project, a file chooser will appear so you can select the directory and project of interest. Once a project has been initialized or opened, a graphical user interface should appear (Figure 2). This is an example of a new project created with four 96 well plates.

Figure 2



4 Images

4.1 Image Formats and Conventions

Currently, *ImageRail* reads grayscale-TIFF images (different images correspond to different filter combinations in the usual manner, thereby enabling multispectral analysis). Before getting started you must make sure that all TIFF images for a given plate are put into their own directory folder. All the images representing each field for all wells and various wavelengths for each individual plate will be mixed together in this directory. (Figure 3). *ImageRail* will parse through and organize the images based on a simple file naming convention.

Figure 3

Name
plate1_B01_2_w530.tif
plate1_B01_2_w685.tif
plate1_B01_3_w460.tif
plate1_B01_3_w530.tif
plate1_B01_3_w685.tif
plate1_B01_4_w460.tif
plate1_B01_4_w530.tif
plate1_B01_4_w685.tif
plate1_B02_1_w460.tif
plate1_B02_1_w530.tif
plate1_B02_1_w685.tif
plate1_B02_2_w460.tif
plate1_B02_2_w530.tif
plate1_B02_2_w685.tif
plate1_B02_3_w460.tif
plate1_B02_3_w530.tif
plate1_B02_3_w685.tif
plate1_B02_4_w460.tif

Each image file name has 4 key properties:

1. Prefix string
2. Well identifier
3. Field number identifier
4. Wavelength identifier

An example file name would look like:

experiment1_A01_1_w488.tif

In this example, the prefix → "experiment1", well identifier → "A01", field number → "1",

and wavelength → "w488". Note that all these identifiers are separated by an underscore ("_").

Also note that the label for an image arising from well B01 must contain a "01;" "B1" will not work.

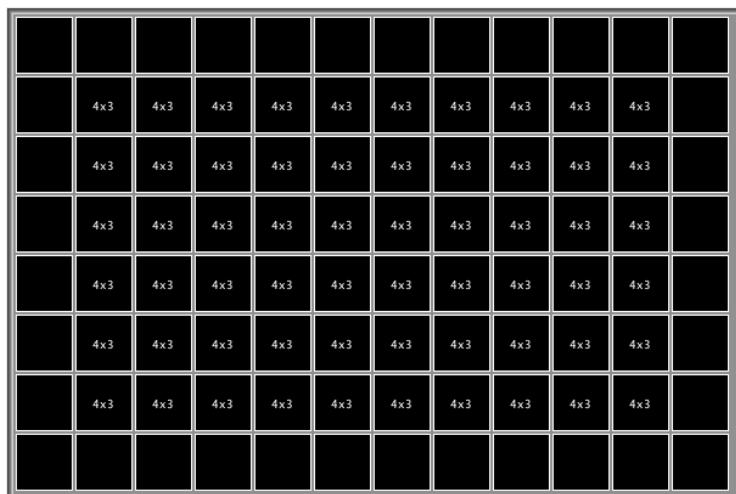
4.2 Loading Images

To begin your analysis, you will first load images into each plate. Please refer to section 4.1 to learn more about current image format requirements. We are actively working to make image formats compliant with Open Microscopy Environment (OME) standards.

Once images conform to the required naming convention and each directory contains images from one plate, drag and drop each directory into the appropriate GUI-plate.

If successfully loaded, the wells of the plate that contain images will now display a NxM indicator, where N represents the number of fields acquired and M represents the number of channels per field (Figure 4). In this example, the outer wells were not imaged.

Figure 4



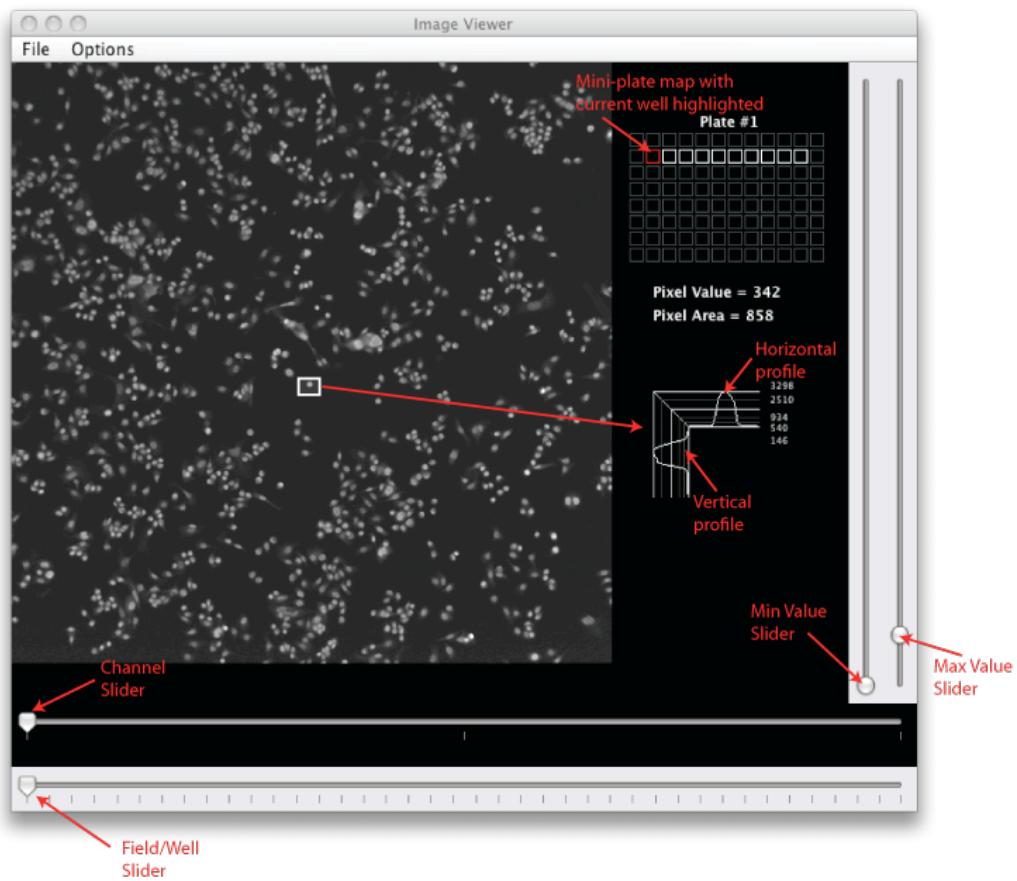
Thus the total number of image files within each well is equal to: NumFields x

NumChannels. In the above example we have 4 fields and 3 channels for a total of 12 separate TIFF images per well.

Note that the original images are **moved** into the *ImageRail* project and not copied, so be sure to make a copy of your images if you want to store them elsewhere as well.

4.3 Viewing Images

Figure 5



4.3.1 Field Viewer

To display the loaded images, click or drag on the plate to select the wells of interest and then either select menu item *Display*→*Display Images* or press *ctrl+d*.

An image browser will open displaying an image, four slider bars and a mini-plate map (Figure 5). NOTE: Mousing over the image will display the pixel intensity on the right margin below the mini-plate.

1. Bottom-most Slider - Changing fields and wells:

The default start image is from the upper left-most of the selected wells in the plate, with its first field and lowest wavelength. The bottom-most slider will first scroll through any other fields of the first well, then move on to the first field of the second well and so forth. The sliding order of the wells is always left to right, top to bottom. The plate map will indicate the current well being viewed, and clicking on a white-highlighted well will skip directly to the first field of that well.

2. Bottom-top Slider - Changing Wavelengths:

Scrolls through the various channels acquired for the currently selected field.

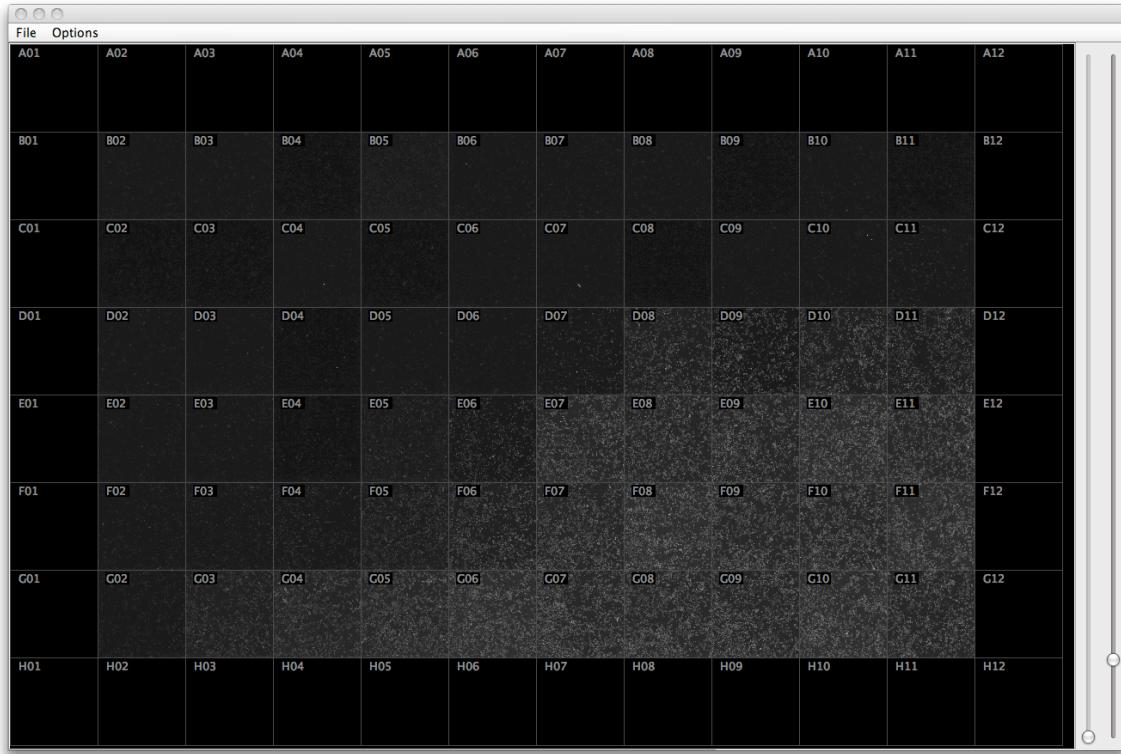
3. Right (1st-left) Slider - Min Value Scaling:

Changes the minimum value of the image; pixels with values lower than this limit are set to 0 (black).

4. Right (2nd-right) Slider - Max Value Scaling:

Changes the maximum value of the image to make it brighter or dimmer; pixels with values higher than this limit are saturated and displayed white.

Figure 6



4.3.2 Montage Viewer

The *montage* button on top of each plate panel will produce a montage image viewer that displays an image from one field from each well (Figure 6). By default, the viewer is initialized with the first wavelength and the first field for each well. The *Options* menu allows selection of other channels and fields. The mouse wheel controls the image zoom. If zoomed, mouse dragging will pan the viewer.

5 Image Processing

Preprocessing

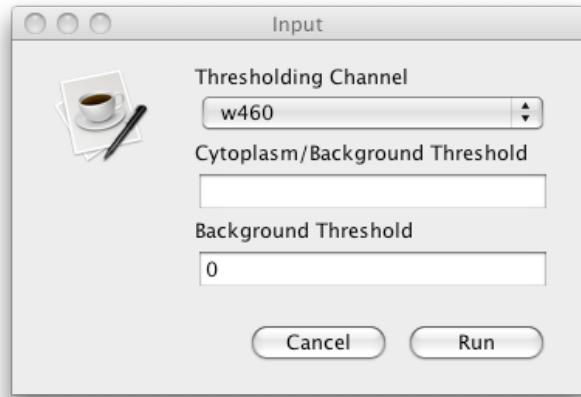
In its default form, *ImageRail* does not perform image pre-processing such as image flat-field calibration, contrast enhancements or convolution filtering. These features as well as other segmentation algorithms can be implemented as plug-ins. When we need preprocessing for our purposes, we use the open-source NIH software program *ImageJ* (<http://rsbweb.nih.gov/ij>).

5.1 Well Means

If you are only interested in population average fluorescent intensities, then there is no need for single-cell segmentation. Well-mean computations can be done quickly. This measurement would be similar to the data acquired from standard fluorescence-based plate readers that specifically ignores image regions without cells, as opposed to performing single-cell analysis and then averaging the data over all cells in the plate.

To begin, highlight the wells in the plate you want to process and use the menu option:
Process → *Well Means*

Figure 7



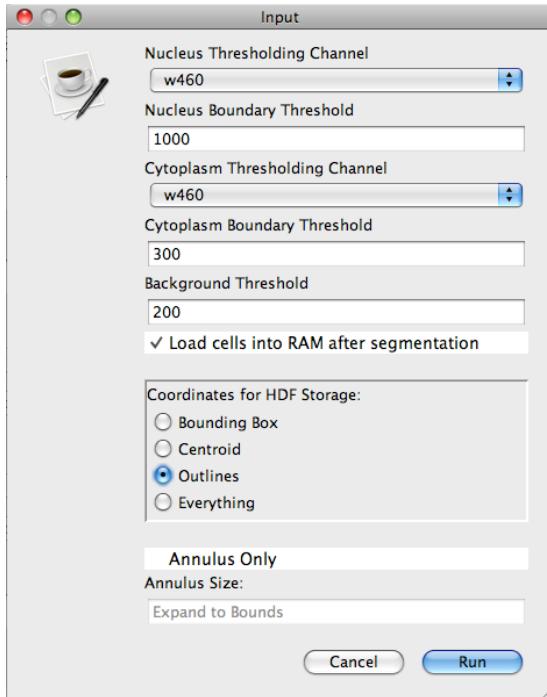
A dialog box will prompt the input of three parameters: 1) channel for cell identification, 2) intensity threshold, which sets a minimum defining where cells exist, and 3) background threshold, which establishes an upper bound for defining background pixels (Figure 7). Refer to section 5.3 to learn how to determine optimal parameters.

5.2 Single-Cell Segmentation

The default *ImageRail* segmentation algorithm is simple yet effective for the 2D tissue cultures for which it was originally designed. When cells grow on top of each other or 3D samples are to be analyzed, more sophisticated segmentation is required. It is currently necessary to do additional programming to integrate new processing routines.

To perform single-cell segmentation of loaded images, use the menu item *Process→Single Cells*

Figure 8



Depending on the version of *ImageRail* or the segmentation algorithm you are using, the dialog window may look different (Figure 8), but in the original version there are three essential user supplied parameters:

1. Nucleus Threshold
2. Cytoplasm Threshold
3. Background Threshold

For both the nucleus and cytoplasm thresholds, the user must specify which channel should be used. NOTE: the channel used to threshold the nucleus can be different from the channel used to threshold the cytoplasm.

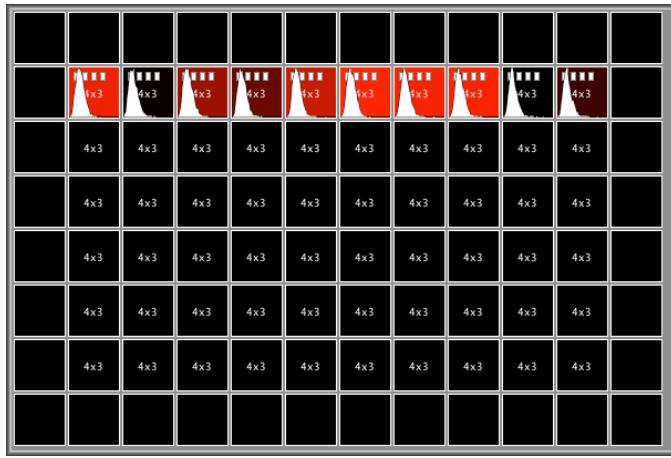
Refer to section 5.3 to learn how to determine optimal parameters.

5.2.1 Load Cells into RAM after Segmentation:

Unless this box is checked, the single-cell data for each processed well will be cleared from

the RAM and only cached to the HDF5 file on the hard drive. If processed for single-cells, the plate will display the number of image fields that have data processed and ready for analysis. These are displayed as icons that look like white document icons within the wells, where each icon represents data for a single field as shown in the top row of the plate (Figure 9).

Figure 9



In this example, the single-cell data has not been loaded. By using the menu option, *Options*→*Load Cells*, or by clicking the “*Load Cells into RAM after Segmentation*” checkbox before processing the wells, the cells will be loaded into RAM and a mini-histogram of all cells will be displayed for the selected feature. Note that there is an option to only load the single-cell data without the pixel coordinates as an alternative to loading both. This will save RAM space if no Data←→Image link is required. See section 6.3 for more details.

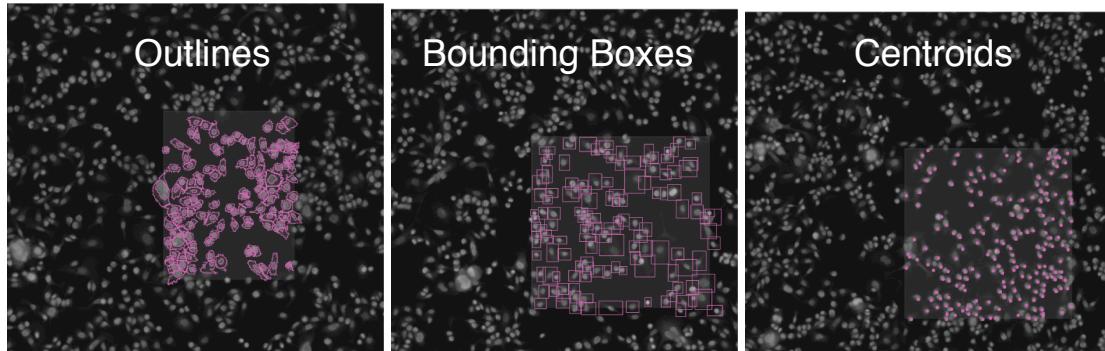
5.2.2 Coordinates for HDF5 Storage:

Once a full segmentation has been performed, all features are computed for each cell using the complete coordinate list. See Supplementary Fig. 4 of *Millard et al.* [1] for a more detailed

description of the “segmentation and feature computation” process.

For long-term storage and data sharing, a complete coordinate list for all cells may not be necessary: in many situations a centroid or bounding box for each cell is sufficient and significantly reduces the amount of data to be stored on the hard drive (Figure 10).

Figure 10

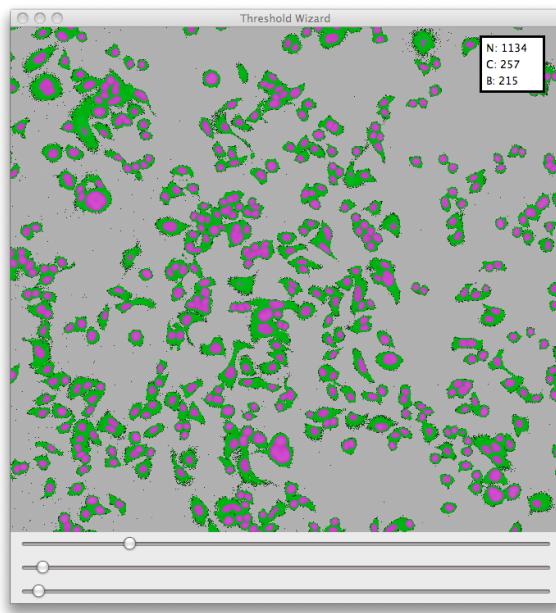


5.3 Parameter Selection

To determine the numerical values for the segmentation parameters, open a sample image of the wells you want to segment. Highlight a region in the image that contains representative cells in the channel that will be used for segmentation and select: *Options→Threshold Wizard*. A separate window will open and display the selected region of the original image and three sliders at the bottom (Figure 11). The top slider changes the nuclear threshold values by coloring pixels greater than the selected value pink. The middle slider does the same for the cytoplasm, coloring pixels greater than the selected threshold green (excluding any enclosed nuclear pixels which remain pink). The bottom slider colors pixels *lower* than the current value white, which represents background pixels. The currently selected threshold values from all sliders are displayed in the upper right corner of the window, where N = nucleus, C = cytoplasm and B = background. Note that even though the green and pink pixel regions may appear contiguous in

the Threshold Wizard, this does not represent the final segmentation result; the segmentation algorithms will subsequently attempt to cut overlapping cells apart with a standard watershed procedure.

Figure 11



In this example, images were acquired using Hoechst stain to identify DNA and a blue whole protein dye to mark the cytoplasm. Note that the stains have drastically different intensity values, such that it is possible to select two different thresholds for the different compartments while using a single fluorescent channel.

In this example, the following parameters were selected:

1. Nucleus = 1000
2. Cytoplasm = 250
3. Background = 200

6 Data Viewing

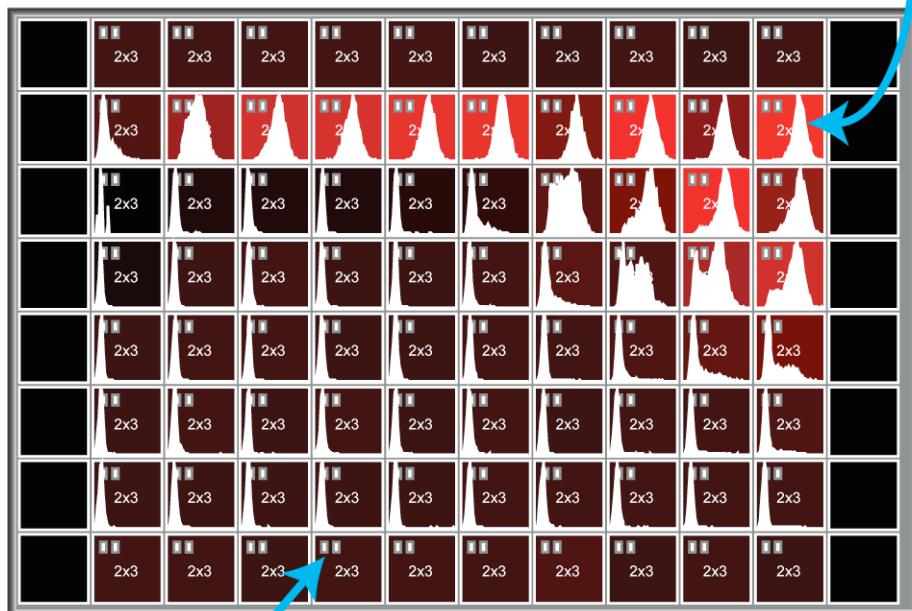
6.1 Plate Graph

The right panel of the main *ImageRail* GUI contains all plates included in the project currently loaded (Figure 12). The graphical plate interface acts both as a controller for actions to be performed on specific wells of specific plates and as a plate-wide data display that includes well heat maps and mini-histograms.

Figure 12

The well background color represents a normalized heat map of the feature selected in the combo-box directly above the plate GUI.

A white histogram represents the single-cell distribution for each well. The y-axis represents number of cells and the x-axis the value of the feature selected in the combo-box above the plate.



A white file icon shows the number of image fields that have single-cell data processed and available to be loaded into the GUI.

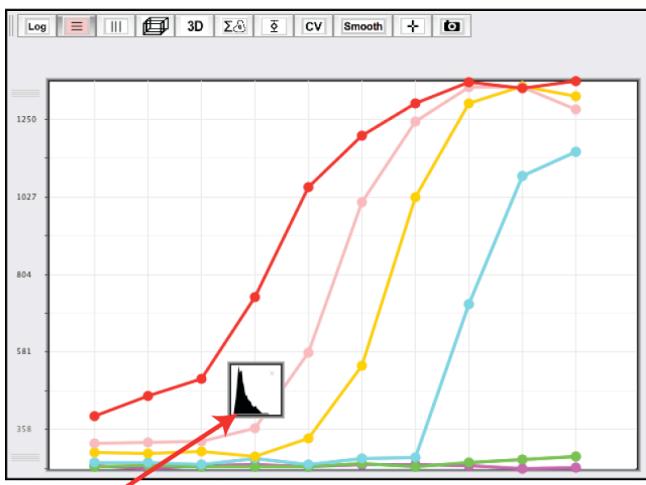
The number of loaded images in each well is indicated by an NxM text string, where N = #Fields and M = #Channels

6.2 Line Plot

Well mean features can be plotted as contiguous well data series in a line plot (Figure 13).

Currently, data series can be constructed based on contiguous wells as rows, columns, or across the same well of multiple plates (trans-plate). Note that if experimental metadata have been entered to denote how each well has been treated, the line plot will attempt to create proper x-axis labels and series legend labels.

Figure 13



Double-clicking on a data point exposes the underlying single-cell population histogram.
NOTE: histograms can be moved via dragging

Toolbar Items:

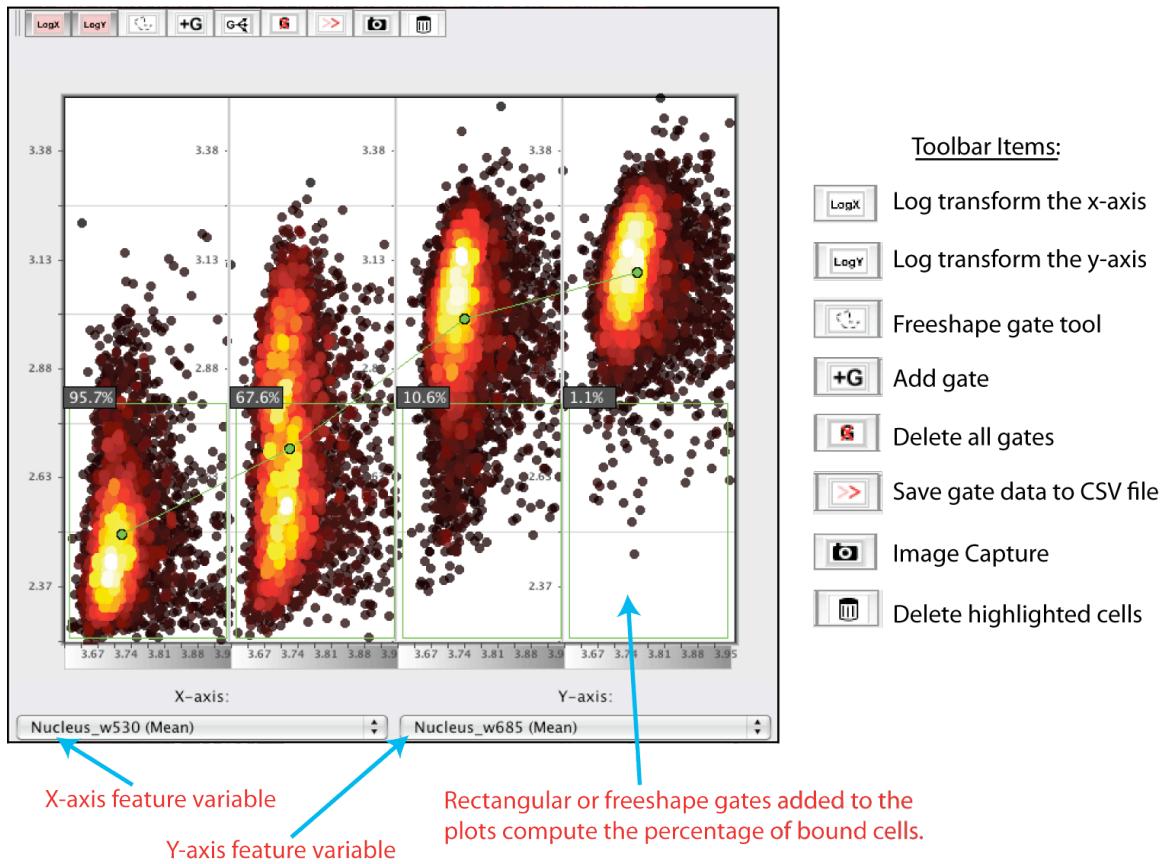
- [Log icon] Log transform the y-axis
- [Row Series icon] Creates row data series
- [Column Series icon] Creates column data series
- [Trans-Plate Series icon] Creates trans-plate data series
- [3D icon] Creates a 3D plot of the data series
- [Sum icon] Plots number of cells in each well
- [Std Dev icon] Plots the population stdev bars
- [Smooth icon] Performs a 3-point smoothing
- [CV icon] Plots the population CV
- [Interpolate icon] Interpolating tracking marker
- [Capture icon] Image capture (JPG, PNG, or SVG)

6.3 Scatter Plot

Single-cell data can be plotted in a 2D scatter plot, where each dot represents one cell with an X-Y coordinate in the plot based on computed features whose identities are selected at the bottom of the plot window (Figure 14).

Figure 14

Cells from four selected wells are plotted in a side-by-side scatter plot for direct comparison. This view also connects the centroids of the distributions by a green line so mean value trends can be visualized.



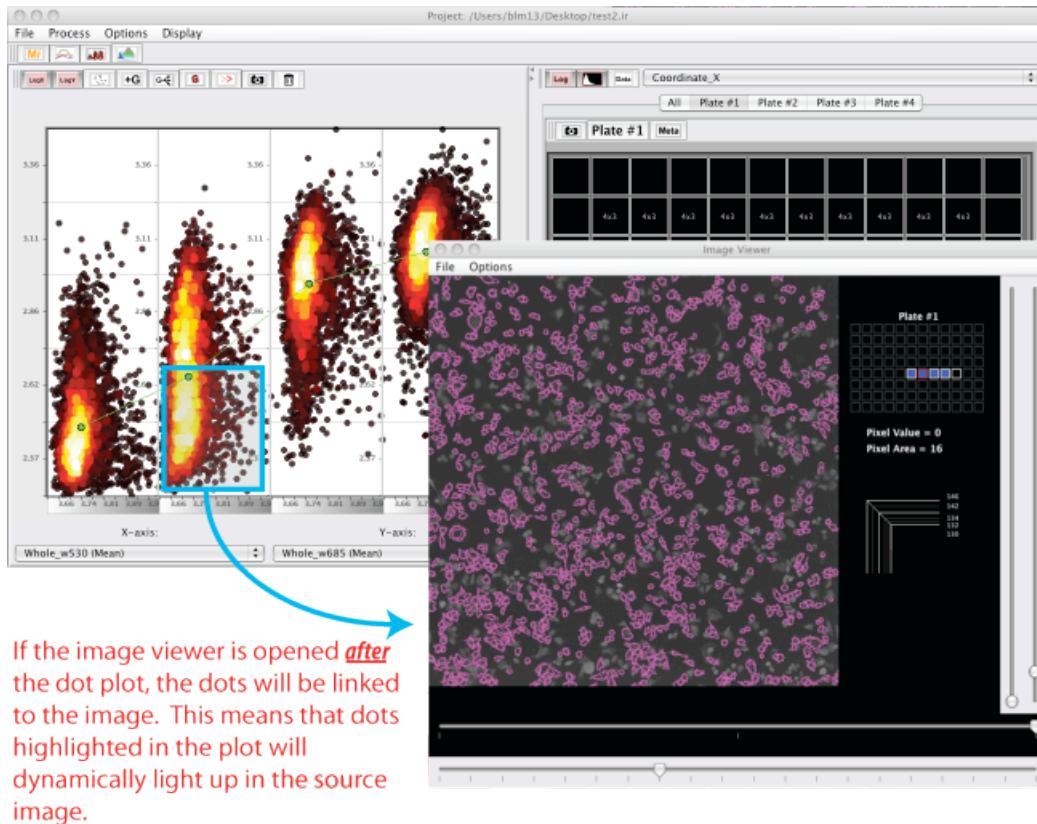
multiple wells are highlighted, then multiple plots will be constructed in a “side-by-side” fashion. The plots are ordered based on well location within the plate: left-to-right, top-to-bottom. A green line connects the centroids of distributions. Note that each well’s scatter plot is a collection of the cells from all image fields within that well.

6.3 Scatter Plot / Image Linkage

The single-cell data within the scatter plot can be directly linked to the source image if the data was loaded via: *Options → Load Cells → Data & Coords*. Once the scatter plot has been

displayed with the appropriate wells highlighted, initiate the image link by opening the image viewer via the menu: *Display* → *Display Images*. Once the image viewer is open, highlighting dots within the scatter plot will dynamically display the corresponding cell within the image (Figure 15).

Figure 15

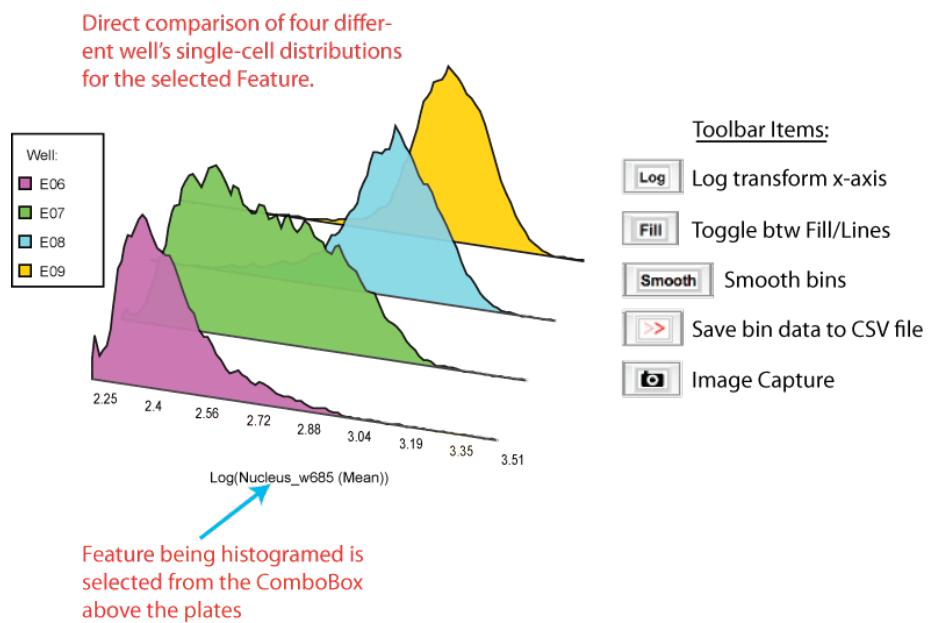


Note that the extent of coordinates stored for this linkage was determined during the segmentation step (see Section 5.2.2). Users can decide to store from as little as a centroid up to all the pixel coordinates that represent the location of each cell. Most often a centroid will suffice for post-segmentation analysis, whereas viewing *outlines* is most useful to confirm proper parameters for segmentation (this is usually done with a test well before switching to centroids for the rest of the plate).

6.4 Histogram Plot

In addition to scatter plots, single-cell data can also be plotted as 1D histograms (Figure 16) of the feature selected in the combo-box at the top of the plate panel. The plots are ordered based on well location within the plate: left-to-right, top-to-bottom. The plot can be translated, zoomed and rotated by holding down *Shift* or *Control* with left-button drag or a right-button mouse drag.

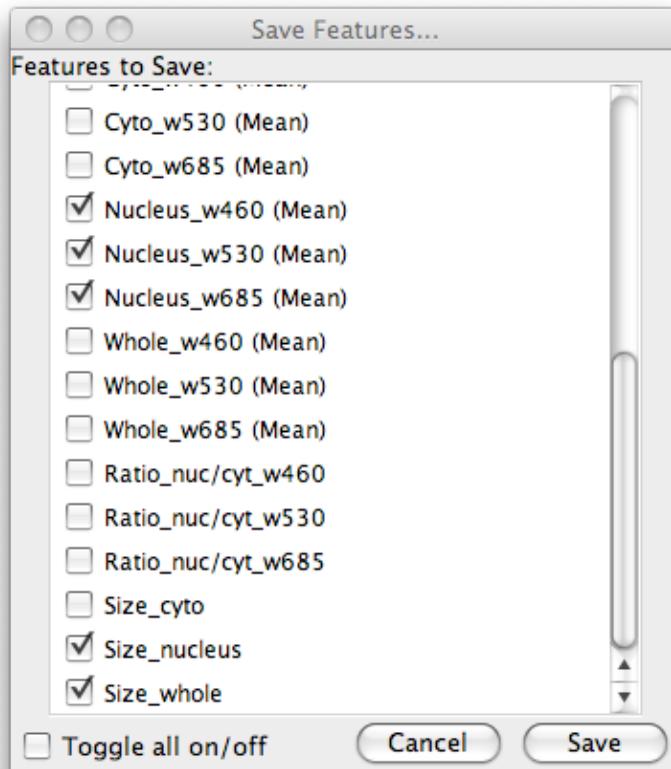
Figure 16



7 Data Export

7.1 CSV – Well Means

Figure 17



Mean value and Standard Deviation data can be exported to a comma-separated value file (CSV) for external analysis in other software programs such as *Excel* and *Spotfire*. To perform the export, use the menu option *File* → *Save as CSV*. A dialog will be displayed that allows the user to select which features to export to CSV (Figure 17). This file will contain blocks of data for each feature organized in an NxM block, where N=number of rows, M=number of columns. The left block represents the well mean feature values and the right block represents the well population standard deviation.

Note: The number of cells for each well will be appended to the bottom of the CSV export

file by default. No selection for cell number is necessary in the dialog window for export.

7.2 MIDAS file formats

MIDAS is a CSV file format capable of encoding experimental data that are coupled to the corresponding experimental design metadata. A complete description is provided in the reference section of publication [2].

7.2.1 Well Means

The MIDAS file format was originally developed for well mean, multiplexed data. To export well means from your plates, highlight the wells you want to export and select the menu item, *File→Save as MIDAS→Well Means*. The dialog box shown at the header of section 7 will allow selection of features that you would like to export. Note that MIDAS export can be done without metadata input.

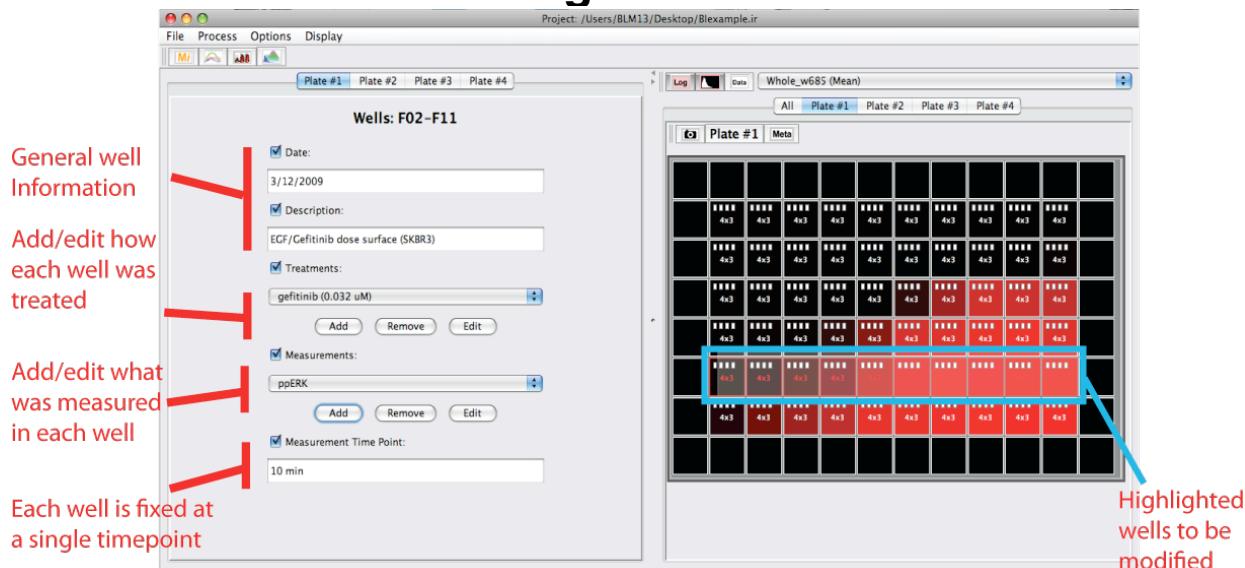
7.2.2 Single-Cells

Although the MIDAS file format was originally developed for well mean values it is capable of storing small amounts of single-cell data. To export single cell data from your plates, highlight the wells you want to export and select the menu item, *File→Save as MIDAS→Single-Cells*. The dialog box shown at the header of section 7 will allow selection of desired features to export. Note that MIDAS export can be done without metadata input.

8 Experimental Design Metadata

To enter experimental design metadata select the yellow “*Mi*” button on the top toolbar of the left panel (Figure 18). *Mi* stands for MIDAS, a file standard for experimental design that we have previously described [2]. First select the wells you wish to annotate in the plate view at right. Then enter the desired metadata in the left MIDAS panel. To modify each field, the checkbox above that field must be selected.

Figure 18



The top two text fields allow for entry of a free-text based Date and Description. Treatments can be added, edited or deleted. A treatment is composed of a compound name, value, units, time of treatment and time units. Measurements for each well can be added to document what each wavelength of the images acquired represents. For example, each fluorescently tagged secondary antibody typically binds a primary antibody specific to a particular protein epitope. Thus fluorescence from each channel is a surrogate read-out for the presence of specific biological components. Note that the addition of a measurement description in the metadata is only for documentation purposes and does not change the exported value

names, which continue to be based on the specific feature that was computed (ex: size_nuclear or Nucleus_w685 (Mean)). Finally, since each well has been fixed at a specific time point, the time of sample acquisition can be added in the bottom text field. Both a time value and time unit can be added (*e.g.*, 10 min), and *ImageRail* will parse the value and the units accordingly.

9 Acknowledgments

We thank Emily Pace, Gerard Ostheimer, Deborah Flusberg, John Albeck, Laura Kleiman and Somponnat Sampattavanich for testing *ImageRail* and providing valuable development insight. This work was supported by National Institute of Health (NIH) grants HG006097, HG005693 and GM68762.

References

- [1] Millard B, Niepel M, Menden M, Muhlich J, Sorger PK. Adaptive informatics for multi-factorial and high content biological data. *Nature Methods* 2011
- [2] Saez-Rodriguez J, Goldsipe A, Muhlich J, Alexopoulos L, Millard B, Lauffenburger D, Sorger PK. Flexible informatics for linking experimental data to mathematical models via DataRail. *Bioinformatics* 2008