

# Bioimage Informatics: Part 2

---

In this experiment, you will learn to use basic bioimage informatics techniques to analyze images of cultured human cells in order to determine whether a targeted mutation in a proto-oncogene affects the cell phenotype.

In this second part, you will use image analysis to count and measure the nuclei of control cells and cells with a mutation in a proto-oncogene.

## 1 OBJECTIVES

### 1.1 EXPERIMENTAL GOAL

Use an image analysis application to quantitatively measure the nuclei count of control A549 cells and A549 cells with a mutant proto-oncogene.

### 1.2 PREREQUISITE SKILLS AND KNOWLEDGE

Basic familiarity with ImageJ.

### 1.3 RESEARCH SKILLS

After this lab, you will have had practice in:

- Using the ImageJ or CellProfiler application to process images of cultured cells and determine the mean number and size of nuclei.
- Using a spreadsheet application to summarize data and perform a t-test.

### 1.4 LEARNING OBJECTIVES

After this lab, you will be able to:

- Explain why cancer can result from genetic mutations.
- Provide an example of a cell-screening assay.
- Describe how bioimage informatics can be used to screen cells for specific phenotypes.

In this experiment, you will learn basic bioimage informatics techniques, and you will use these techniques to begin analyzing images of cultured human cells to determine whether a mutation in a proto-oncogene can affect cell proliferation in culture.

## 2 PRE-EXPERIMENT

### 2.1 CANCER AS A GENETIC DISEASE

Human cells have carefully controlled molecular and biochemical mechanisms to keep the cells in the body growing and dividing at an appropriate pace for the benefit of the organism. In cancer, this carefully balanced system of growth and restraint goes awry.

Cancer is caused by certain mutations in the DNA of cells. Some mutations are inherited, but most occur during a person's lifetime as a result of random errors in replication. Environmental factors that damage DNA, such as smoking and sunlight, can also cause mutations to occur.

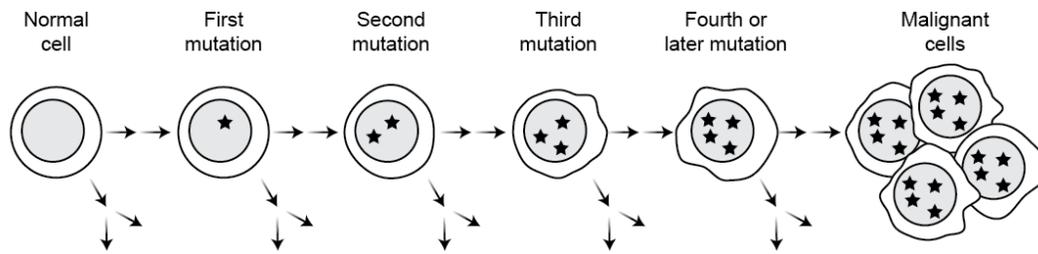


Fig. 1: Schematic of cancer development. Cells accumulate mutations as they divide. Mutations that are most advantageous for cell growth and survival are passed on to daughter cells, which, in turn, acquire further mutations and may eventually become malignant cancer cells. (The arrows represent multiple cell divisions.)

As a single cell in the body accumulates mutations, one of those mutations could provide a survival or a growth advantage to the cell, causing the cell to divide at a faster-than-normal pace or to not die. The resulting daughter cells with that mutation, which are dividing quickly, are more likely to accumulate additional mutations. Additional mutations that affect cell division may cause those cells to divide even faster. Eventually, a cell may acquire enough mutations that it starts to grow and divide uncontrollably (Fig. 1).



**Before continuing**, watch the following 8-minute video clip on large-scale efforts to identify genetic mutations associated with human cancers: <http://www.hhmi.org/biointeractive/cancer-genetic-disease-video-highlights>

As you learned in the video, researchers have thus far identified about 140 different genes that, when mutated, can lead to the development of cancer. Most cancers contain mutations in two to eight of these genes. The 140 or so genes that drive the development of cancer fall into two categories that describe their effects on cell division. One category consists of oncogenes, which are mutated genes whose protein products cause cells to divide faster. The normal versions of oncogenes are called proto-oncogenes; they code for proteins that stimulate normal growth and division. The second category consists of tumor suppressor genes, which encode proteins that normally put the brakes on cell division and, when mutated, have the effect of taking the brakes off.

Oncogenes and tumor suppressor genes affect several different cellular processes, which can be grouped into cell growth and survival (genes involved in the cell cycle), cell fate (genes involved in cellular differentiation), and genome maintenance (mostly DNA repair genes).

Each of these cancer-associated genes codes for a protein that has a normal function within cells, and this normal function does not lead to the development of cancer. However, a mutation in the gene may prevent it from coding for a functional protein, or it may cause it to code for a protein that has an altered function. For example, the mutation may lead to reduced or absent expression of a tumor suppressor gene, or it may lead to excessive activity of a protein that stimulates cell growth and division.

## 2.2 EPIDERMAL GROWTH FACTOR RECEPTOR (EGFR)

The epidermal growth factor receptor (EGFR) is a cell-surface receptor protein that is activated by binding to certain growth factors, including epidermal growth factor and transforming growth factor  $\alpha$  (TGF $\alpha$ ). EGFR is part of the ErbB family of receptors, a group of closely-related receptors that are receptor tyrosine kinases. EGFR helps cells respond to signals that regulate normal growth and division. When EGFR is activated by binding to a growth factor, it autophosphorylates some of the tyrosines on the intracellular side of the protein. That is, EGFR is a tyrosine kinase that, upon binding to a growth factor, activates itself. This in turn activates signal transduction cascades that increase cell division and cause changes in cell phenotypes such as cell migration and adhesion.

A number of cancers, including some lung cancers, are associated with genetic mutations that lead to excessive activity of EGFR or increased numbers of EGFR proteins (overexpression), both of which can cause uncontrolled cell division. Therefore, EGFR is an example of a proto-oncogene that, when mutated, is an oncogene. Important anticancer drugs have been developed that interfere with the activation or signaling of EGFR. However, each anticancer drug is only effective for certain patients, and some of the difference in effectiveness between patients depends on the specific genetic mutation(s) in each patient that caused the EGFR to be overactive or overexpressed.

### 2.3 USING CELL CULTURE AND MOLECULAR BIOLOGY TO STUDY ONCOGENES

Identifying the genes and genetic mutations associated with human cancers is essential in understanding the role these genes and mutations play in causing cells to become cancerous. Molecular biology techniques give us the ability to insert specific mutations into one or more genes of cells in vitro. The goal is to identify specific mutations that cause differences in the phenotype of a cell that are potentially consistent with cancer cells. When we identify such a mutation, this gives us an experimental model with which to study the functions of that gene mutation in cancer, and it gives us an in vitro model to test the effectiveness of drugs that might control or even selectively kill cancer cells with that mutation. In these kinds of experiments, one is said to be “screening” for putative oncogene mutations.

### 2.4 BIOIMAGE INFORMATICS

The image analysis techniques you utilized in the previous exercise can be used to screen cells from patients for altered phenotypes. However, as described above, at least 140 genes are associated with cancers in humans, and it turns out that each of those genes has many sites at which mutations can lead to altered gene function. Therefore, conducting an oncogene screen requires assessing changes in cell phenotype for at least 140 genes, each with potentially dozens of different mutations, and in many different types of cell lines. Furthermore, any good experimental design would include replicates (repeats of the same experiment to ensure the same results are obtained). Consequently, a screen will typically involve measuring dozens of phenotypes in hundreds of thousands of cells. Although this type of analysis could be performed using the techniques you learned in the previous exercise, it would take tremendous effort and time for the researcher to measure each cell individually.

To address this need, researchers have developed computerized image analysis approaches that allow automated measurements of cellular phenotypes from thousands of images. This screening approach can yield millions of data points from a single experiment, which is a “big data” or “informatics” approach to analyzing gene function. Therefore, it is sometimes referred to as bioimage informatics.

In today’s exercise, you will use an image analysis application to do a simple screen of several images of human A549 cells to determine whether a specific genetic mutation in the gene for EGFR causes phenotypic changes consistent with cancer. You will have three sets of images representing three experimental groups: control A549 cells (expressing the normal EGFR gene), A549 cells with a mutation in the gene for EGFR (mutant allele 1), and A549 cells with a different mutation in the gene for EGFR (mutant allele 2). You will analyze the control cells and cells with one of the mutations, as assigned by your instructor.

### 3 LABORATORY MANUAL: IMAGEJ OPTION

#### 3.1 MATERIALS CHECK OFF LIST

Each small group of (2-3) will have:

- 1 laptop computer with ImageJ loaded.
- Digital images of DNA-labeled A549 cells for control cells and two EGFR mutant alleles, all with 8 replicates and 9 sites per replicate.

#### 3.2 SAFETY AND WASTE DISPOSAL PROTOCOLS

No protective gear is required for this lab. Do not eat, drink, or apply anything to the skin while in this laboratory. Please leave the desktops clean.

#### 3.3 EXPERIMENTAL PROCEDURE

##### 3.3.1 Make Predictions

You will use image analysis approaches to measure the cell count for control cells and cells with mutations in the EGFR proto-oncogene.

- Q1. Predict whether the mutation in EGFR will increase, decrease, or have no effect on the number of cells and the average nuclear size.

##### 3.3.2 Source Images

At the beginning of the session your instructor will tell you whether the images are already on your computer. If they are not, then download and extract the zipped folder “Mutation\_Screen\_Images” in the resources folder on the course home page.

These are all monochrome images of A549 cells acquired with a low-power objective optimized for fluorescence microscopy of many cells and samples at a time. Prior to the cells being imaged, they were labeled with five fluorescent dyes: one that labels DNA; one that labels RNA; one that labels endoplasmic reticulum (ER); one that labels mitochondria; and one that labels actin, the Golgi apparatus, and the plasma membrane (AGP). You will only be using the images of the dye that labels DNA.

Your image set has a “Control” group and two experimental groups with EGFR mutations. For each group, there are eight replicate multiwell tissue culture plates (replicates 1-8). Within each replicate multiwell plate, images were acquired at nine sites (s01 to s09).

##### 3.3.3 Create an Output Folder

1. Create a new folder called “Analysis Output” on the desktop:
  - a. On the desktop of the computer, right click on the desktop, select “New”, then “Folder”.
  - b. Rename the folder “Analysis Output”.

##### 3.3.4 Open an Image and Calibrate the Scale

The instructions in this laboratory exercise will refer to ImageJ, but note that your laboratory computer may instead have a program installed called FIJI. This is just Image J with many additional preloaded plug-ins and modules that extend its functionality. You may use ImageJ or FIJI interchangeably for this exercise.

2. Open the ImageJ application.

Recall from the preceding exercise that to undo the most-recent edit or image manipulation, select **Edit>Undo**. Note that *only one back-step is possible*. To convert all changes back to the most recently saved version, select **File>Revert**.

3. Open the control, replicate 1, site 1 image file: r01\_s01\_DNA.tif.
  - a. Recall that to open an image file, select File > Open from the menu bar, or simply drag the image file onto the ImageJ menu bar. Navigate to the following folder: Mutation\_Screen\_Images > Control > replicate\_1 and open r01\_s01\_DNA.
  - b. Drag the window edges or corners to resize it on your screen, if desired.
4. Adjust the zoom level.
  - a. As noted in the previous exercise, the zoom level of the image is shown in the image title field at the top of the image window. For example, if the zoom level is 150%, the title field will show “r1\_s01\_DNA.tif (150%)” (Fig. 2).
  - b. To zoom in, press the plus (+) or up-arrow (↑) keys, and to zoom out press the minus (-) or down-arrow (↓) keys.
  - c. The image window will resize if the displayed image is smaller than the window, but if the image cannot fit within the window, you will see the *Zoom indicator* in the upper left corner that shows what portion of the image is currently displayed (Purple boxes in Fig. 2).
  - d. Adjust the zoom level so that you can see the entire image in the window (i.e., no Zoom indicator is displayed).

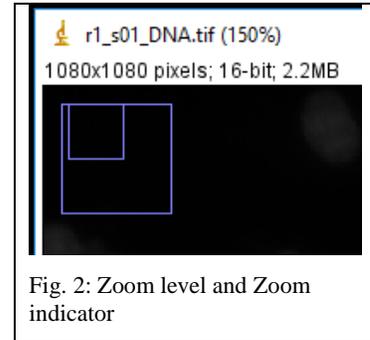


Fig. 2: Zoom level and Zoom indicator

- Q2. Discuss the distribution of nuclei across the field of view. Are the nuclei randomly distributed? Are the nuclei separated from each other, or are any overlapping or touching? Do the nuclei all have the same shape and brightness?
5. Set the distance scale.
    - a. In the Menu Bar select “Analyze” and then “Set Scale...” For the low-magnification images you are looking at in this exercise, each pixel is 0.66 microns across. Thus, for *Distance in pixels* enter 1, for *Known distance* enter 0.66, and for *Unit of length* enter “micron”. Then check the Global box and click OK.

### 3.3.5 Count and Measure the Nuclei

You will now use ImageJ to automatically count and measure the nuclei in an image.

6. Estimate the size range of nuclei.
 

The nuclei in the image have a distribution of sizes. In order for ImageJ to automatically identify nuclei in the image, you will need to determine the upper and lower range of sizes that you will accept.

  - a. Zoom in the image to 150% or 200%.
  - b. Select the Scroll tool () and then use the mouse/trackpad to scroll (pan) through the image to find a region with a representative sample of nuclei.
  - c. Find one of the smaller “particles” in the window that you are reasonably confident is a nucleus and is not noise, dirt, or some other non-nuclear material.
  - d. Use the Polygon selection tool or Freehand selection tool to outline the nucleus, then press Ctrl+M to record the area.
  - e. Repeat this for a few more of the smaller nuclei.
  - f. Repeat the preceding two steps for the largest particles that are not clusters of overlapping or touching nuclei or some other non-nuclear material.

Q3. What are the minimum and maximum areas of the nuclei you sampled?

7. Apply a Gaussian blur.
  - a. To smooth the edges of the nuclei and reduce the noise, select Process > Filters > Gaussian Blur...

- b. Set the Sigma (radius) to 2.00.
  - c. Click OK.
8. Convert the image to black and white.
  - a. Select Process > Binary > Make Binary.
  - b. The image should now show black nuclei against a white background. If it is the opposite from this (i.e., white nuclei against a black background), select Edit > Invert.
9. Count and measure the particles.
 

The binary image you have just created has black “particles” against a white background. You have determined that most of these particles represent nuclei. The Analyze Particles function allows ImageJ to count the number of particles in the image that meet your criteria for size.

  - a. Select Analyze > Analyze Particles...
  - b. Set the lower value of the size range (in square microns) to 100, which should be consistent with the minimum area of the sample you measured above, and set the upper range to infinity (i.e., “100-infinity”).
  - c. Leave Circularity at the default setting (0.00-1.00).
  - d. In the Show dropdown menu, select Overlay Masks. This displays filled outlines over the particles in the image.
  - e. Check the following options: Display results; Clear results; Summarize, Exclude on edges; and Include holes.
  - f. Click OK.
  - g. After ImageJ processes the image, you will see light blue masks over the counted particles in the image window, each of which has its count inside, and you should see a Results window and a Summary window.
    - i. The “Results” window lists each object, its area, and the mean, minimum and maximum darkness (pixel intensity) of the object. These darkness values are irrelevant because the image was converted to binary, and therefore all the particles are black (intensity of 255).
    - ii. The “Summary” window list the title of the image, the particle count, the total particle area (sum of all particles in the image), the average particle area (“size”), the percent of the total image area that is particles, and the mean darkness (pixel intensity) of the particles. As in the Results window, this last item is irrelevant because the image is binary.
  - h. both of which will show that slightly more than 200 particles were counted and measured. The black particles in the image were not included.

Q4. What are likely reasons why the various black particles were not included?

### 3.3.6 Apply Cell Segmentation in a New Image

You will now examine an image with many touching nuclei and use segmentation to improve the automated analysis.

10. Close the r1\_s01\_DNA window.
11. Open image r1\_s02\_DNA.tif.
12. Apply a Gaussian blur, as before, using Process > Filters > Gaussian Blur..., with a Sigma of 2.00.
13. Convert the image to black and white using Process > Binary > Make Binary.
14. Count the particles using Analyze > Analyze Particles with the same settings as before (these should all already be set). When ImageJ asks if you want to save the previous measurements, select the “No” option. This will clear the data from the Results window but it will not remove the summary data from the Summary window. The new data will display in the Results window, and the Summary window will now update to include a second row with the new measurements. Both windows should show that about 320 particles were counted and measured.

Q5. Why did the automated analysis not count and measure the nuclei as accurately as in the previous image?

A number of computational methods have been developed to cut or “segment” touching particles. In this exercise you will use the *watershed segmentation* method.

15. Select Process > Binary > Watershed.
  - a. Note ImageJ now creates a white line between many of the touching nuclei.
16. Count and measure the particles again using Analyze > Analyze Particles.
  - a. You should now have about 430 particles.

Q6. Are there still black particles that you think should have been included?

Q7. What could you change to include more of the particles?

17. Count and measure the particles again using Analyze > Analyze Particles, but reduce the minimum size from 100 square microns to 50 square microns. Leave the upper limit as infinity. As before, select “No” when ImageJ asks if you want to save the previous measurements.
  - a. You should now have about 470 particles.

With the information you have available, it’s not clear whether this is a more accurate analysis. On the one hand, reducing the minimum size allowed some nuclei to be included that were left out before. On the other hand, it appears that some single nuclei were segmented into multiple particles, which resulted in an over-count and under-measurement of the area. If you don’t have additional information available, the size and circularity ranges are more a matter of judgment combined with trial-and-error. For the remainder of this exercise, you’ll use the current settings (“50-infinity” range, “0.00-1.00” circularity) for the remaining nuclei counts. Although the absolute counts may not be accurate, it will let you make reasonable comparisons between images.

### 3.3.7 Use Stacks to Automatically Analyze Multiple Images

Now that you’ve established settings to use for counting the nuclei, you can use a *stack* of images to further automate the process. A stack in ImageJ is two or more images that are displayed in a single window. The images that make up a stack are called *slices*. Many operations in ImageJ, including converting to binary images, watershed segmentation, and particle counting, can be applied to all slices in a stack.

18. Create a stack of the nine sites in the control, replicate 1 image set.
  - a. Close the r1\_s02\_DNA window, the Summary window, and the Results window (do not save the measurements).
  - b. Open all nine control, replicate 1, DNA images (r1\_s01\_DNA, r1\_s02\_DNA, ... r1\_s09\_DNA).
  - c. Select Image > Stacks > Images to Stack
    - i. It is not necessary to name the stack because you will not be saving it.
    - ii. Note that the resulting stack window has a scroll bar across the bottom that lets you cycle through the nine slices.
19. Apply a Gaussian blur, as before, using Process > Filters > Gaussian Blur..., with a Sigma of 2.00.
  - a. Select “Yes” when you are asked whether to process all 9 images.
20. Convert the images in the stack to binary
  - a. Select Process > Binary > Make Binary
  - b. For Method, use the default setting.
  - c. For Background, select Dark.
  - d. Make sure “Calculate threshold for each image” is checked.
  - e. The window will preview the calculated binary conversion, with red indicating pixels that are above the calculated threshold.
  - f. Click OK.

21. Segment the nuclei
  - a. Select Process > Binary > Watershed.
  - b. Select “Yes” when you are asked whether to process all 9 images.
  - c. Use the horizontal scroll bar to examine the results. Note that the order of images will not necessarily match the order of sites, but the original file names are preserved in the title bar of each window.
22. Count the nuclei
  - a. Select Analyze > Analyze Particles
  - b. Confirm that the settings are the same as before.
  - c. Click OK.
  - d. Select Yes when you are asked “Process all 9 images?”
  - e. Select “No” if you are asked whether to save the measurements. This refers to the measurements from the prior analysis.
  - f. A “Summary of Stack” window will be created that lists the title of each image (slice) and the summary results for each image. The Results window will show the area for every particle included in the stack (i.e., all of the images).
23. Save the summary data.
  - a. Click on the Summary of Stack window and select File > Save As..., give the file a descriptive name (such as “Control\_R1.xls”), and save it to the “Analysis Output” folder you created earlier on the desktop.

### 3.3.8 Analyze Additional Replicates of the Control Treatment

24. Close the stack.
25. Repeat the steps in the preceding section for at least two additional replicates of the control treatment. These are in the folders named replicate\_2, replicate\_3, etc. up through replicate\_8. Your instructor may indicate the number of additional control replicates you should analyze. In each case, save the “summary of stack” file with a unique name that indicates the replicate number.

### 3.3.9 Analyze the Images of the Mutant Allele

Now that you have determined the nuclei count on the control cells, you will perform the same analysis on the images of cells with a mutant allele.

26. Close any open image and results windows in ImageJ.
27. Create a stack of the images from the nine sites in the **replicate\_1** image set in either the **EGFR\_Mutant\_1** folder or **EGFR\_Mutant\_2** folder, as assigned by your instructor.
28. Apply a Gaussian blur, as before.
29. Convert the stack to binary images.
30. Segment the nuclei using watershed segmentation.
31. Count the nuclei using Analyze Particles, with the same settings as you used for the control images.
32. Use the horizontal scroll bar to check the results of the analysis.
33. Save the “summary of stack” results in the “Analysis Output” folder with a descriptive name (such as “Mutant\_1\_r1.xls”).
34. Repeat the preceding steps for at least two additional replicates of the mutant allele treatment. Your instructor may indicate the number of additional control replicates you should analyze.
35. When you are done performing the image analyses, close the ImageJ application.

## 3.4 STATISTICAL ANALYSIS

You will now determine whether the mutation in the EGFR gene significantly affected the nuclei number or nuclear area. To do this, you will use a t-test to calculate the probability that any difference in the mean number or mean area between the control and mutant treatments is due to chance.

- Q8. For the purposes of statistical analysis, can each of the nine sites (images) within each replicate be considered a replicate? That is, if you determined the mean nuclei count for nine images for each of five replicates, do you have 45 replicate measurements or only five replicate measurements? Explain your answer.

### 3.4.1 Determine the Replicate Means

Your output data are in separate files for each replicate (stack). For example, if you analyzed four replicates each of the control and mutant treatments, you should have eight output files. You first need to determine the mean count and mean area for each replicate.

36. Open a spreadsheet application (e.g., Microsoft Excel or Google Sheets), create a new spreadsheet.
37. In the first row, enter the following as headings in columns A-E: Replicate, Control Count, Control Area, Mutant Count, Mutant Area.
38. In the Replicate column (A), enter 1 and cell A2, enter 2 in cell A3, and continue until you have created a row for each replicate you analyzed.
39. Save this with the filename “Mutation\_Screen\_Summary” in the Analysis Output folder on the desktop, but keep the file open.
40. Open the output file for the first control replicate (e.g., Control\_R1.xls).
41. Calculate the means for the Count column and the Average Size column. If you are unsure how to do this within the spreadsheet application, review [Spreadsheet Tutorial 1](#) (Formulae, Functions and Averages) on the HHMI BioInteractive [Spreadsheet Data Analysis Tutorials](#) site.
42. Copy and paste the values for the mean count and mean area into the appropriate cells for count and area in the second row of the Mutation\_Screen\_Summary spreadsheet. Note that to make sure you paste the value and not the formula, in Sheets use Paste Special > Paste Values Only, and in Excel use Paste Special > Values.
43. Repeat the preceding two steps for each of the remaining control replicates and then for each of the mutant replicates.
44. When you are done, save your Mutation\_Screen\_Summary spreadsheet, but keep it open. Close any other open spreadsheet windows.

### 3.4.2 Test for Differences Between Means

45. Perform a two-tailed t-test for count data in the Mutation\_Screen\_Summary spreadsheet. If you are unsure how to do this within the spreadsheet application, review [Spreadsheet Tutorial 4](#) (t-test) on the HHMI BioInteractive [Spreadsheet Data Analysis Tutorials](#) site.
- Q9. Did you observe a significant effect of the EGFR mutation on the mean nuclei count? Did the result match your prediction?
46. Repeat the preceding steps for the nuclear area data.
- Q10. Did you observe a significant effect of the EGFR mutation on the mean nuclei area? Did the result match your prediction?
47. When you are done, save your Mutation\_Screen\_Summary spreadsheet.

## 4 LABORATORY MANUAL: CELLPROFILER OPTION

### 4.1 MATERIALS CHECK OFF LIST

Each small group of (2-3) will have:

- 1 laptop computer with CellProfiler loaded.
- Digital images of DNA-labeled A549 cells, with 8 replicates and 9 sites per replicate, for two treatment groups: control, and EGFR mutant 1.

### 4.2 SAFETY AND WASTE DISPOSAL PROTOCOLS

No protective gear is required for this lab. Do not eat, drink, or apply anything to the skin while in this laboratory. Please leave the desktops clean.

### 4.3 EXPERIMENTAL PROCEDURE

#### 4.3.1 Make Predictions

You will use image analysis approaches to measure the cell count, cell area, and nuclear area for control cells and cells with mutations in a proto-oncogene.

Q11. Predict whether the mutation in EGFR will increase, decrease, or have no effect on the number of cells, the average cell size (area) and the average nucleus size (area).

#### 4.3.2 Source Images

At the beginning of the session your instructor will tell you whether the images are already on your computer. If they are not, then download and extract the zipped folder “Mutation\_Screen\_Images” in the resources folder on the course home page.

These are all monochrome images of A549 cells acquired with a low-power objective optimized for fluorescence microscopy of many cells and samples at a time. Prior to the cells being imaged, they were labeled with five fluorescent dyes: one that labels DNA; one that labels RNA; one that labels endoplasmic reticulum (ER); one that labels mitochondria; and one that labels actin, the Golgi apparatus, and the plasma membrane (AGP). You will only be using the images of the dye that labels DNA.

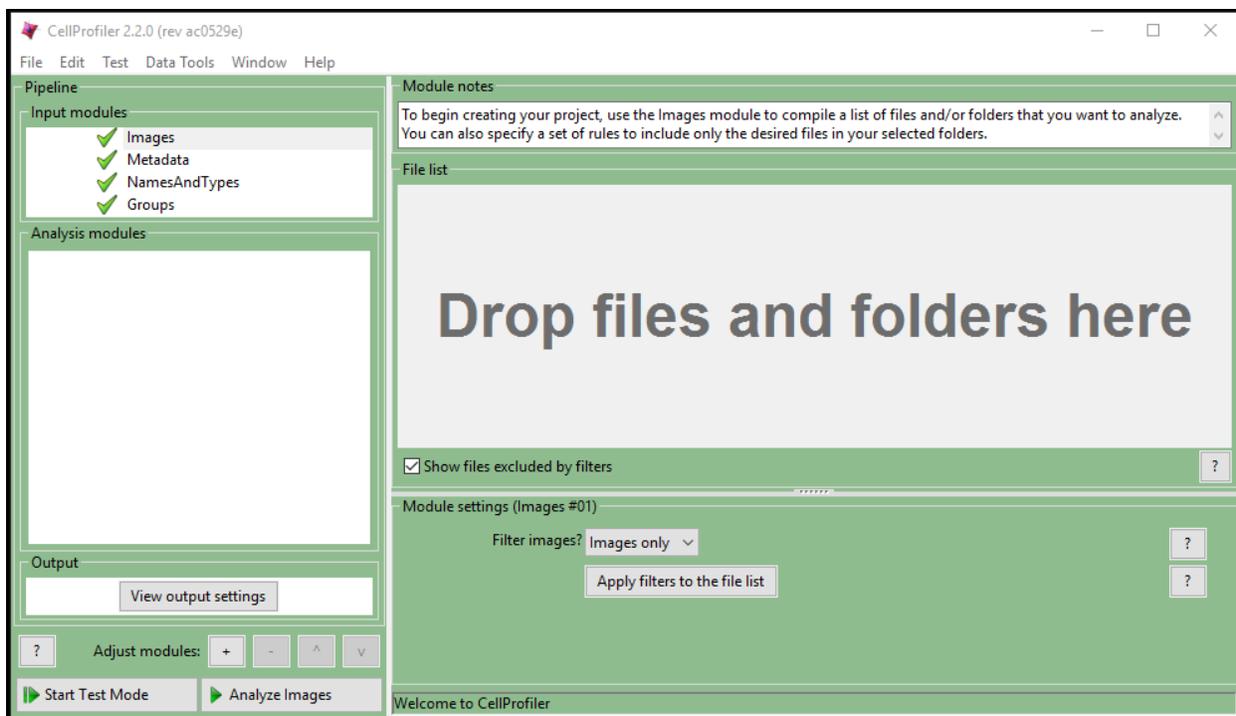
Your image set has two experimental groups: “Control”, and “EGFR Mutant 1” allele 1. For each experimental group, there are eight replicate multiwell tissue culture plates (replicates 1-8). Within each replicate multiwell plate, images were acquired at nine sites (s01 to s09).

#### 4.3.3 Create an Output Folder

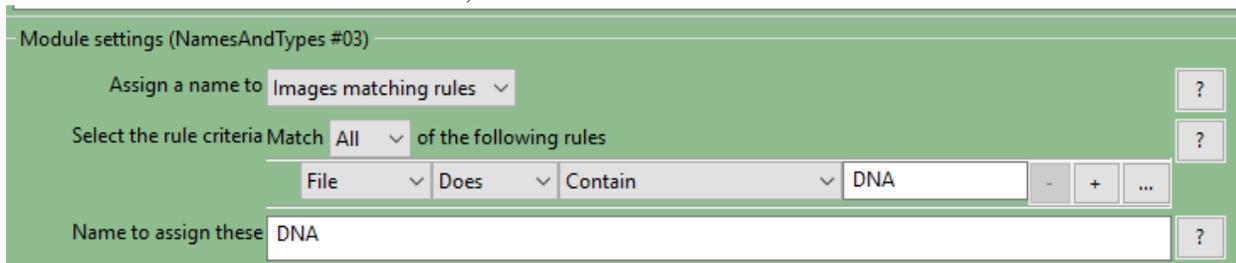
48. Create a new folder called “Analysis Output” on the desktop:
  - a. On the desktop of the computer, right click on the desktop, select “New”, then “Folder”.
  - b. Rename the folder “Analysis Output”.

#### 4.3.4 Set Up CellProfiler to Recognize Nuclei

49. Open CellProfiler.
50. You will see multiple windows: the “CellProfiler.exe” window (if you are using a Windows computer, but not if you are using an iOS or Linux computer), which you can minimize but not close; the “Welcome to CellProfiler” window, which you won’t need today and can close; and the main CellProfiler window, shown below.



51. Upload the Control Images into CellProfiler:
  - a. Locate the Control folder inside the Mutation\_Screen\_Images folder.
  - b. Drag the Control folder into the File list box (indicated by “Drop files and folders here”).
52. Set the Output folder:
  - a. Click on the “View output settings” button in the Output window on the left side of CellProfiler.
  - b. For the Default Output Folder, click on the folder icon () , then navigate and select the “Analysis Output” folder you created on the desktop.
53. You will now tell CellProfiler which image files are the DNA channel image files:
  - a. Click the “NamesAndTypes” module, which is inside the “Input modules” box on the left side.
  - b. In the green “Module notes” box on the right side, select the “Assign a name to” drop-down menu and select “Images matching rules”.
  - c. Set the “Select rule criteria Match” to select “File Does Contain” and enter “DNA” in the match criteria text box, and then enter “DNA” as the name to assign the image files that match the selection criteria, as shown below

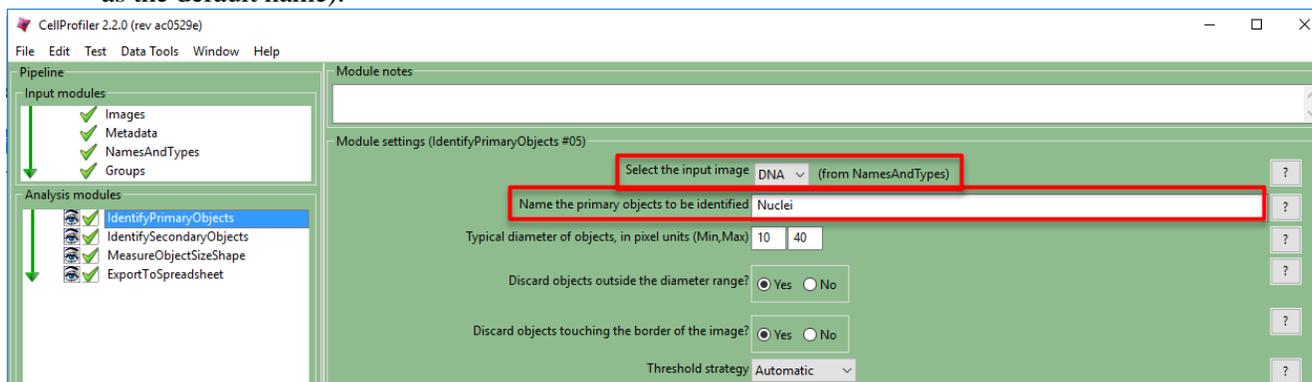


- d. Click Update. You should now see a column of the control DNA channel image files under the heading DNA. With this set up properly, you will be able to specify all of your individual DNA channel images simply by referring to “DNA”.

CellProfiler contains a large number of analysis modules. To run an automated analysis, you must create a “pipeline”, which is a series of analysis modules that together perform your desired task. The analysis pipeline begins with the four input modules you have just been working with, which load and name the image files, and then proceed to the analysis modules. You will now begin adding analysis modules to create a complete pipeline.

In this exercise you will use four analysis modules: 1) Identify Primary Objects, 2) Identify Secondary Objects, 3) Measure Object Size Shape, and 4) Export to Spreadsheet.

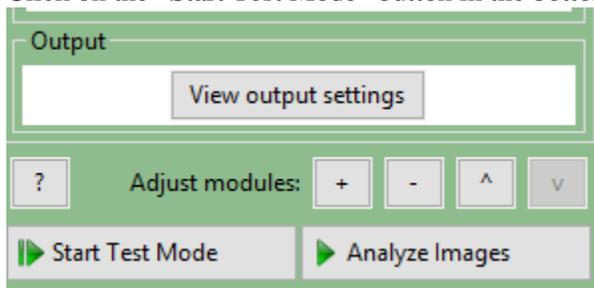
54. You will now provide CellProfiler with instructions for how to recognize the nuclei in the control image data set:
  - a. In the menu bar, select “Edit”, then “Add Module >”, then “Object Processing >”, then “IdentifyPrimaryObjects”.
  - b. In the Module notes window on the right, find the “Select the input image” field and select “DNA” from the drop-down menu. This tells CellProfiler to use the DNA images (which you identified in the Names and Types input module) for this module. In your experiment, you have 72 images, but this same pipeline could be applied to millions of images, especially when using cluster-based or cloud-based computing.
  - c. In the “Name the primary objects to be identified” field, enter “Nuclei” (it may already be there as the default name).



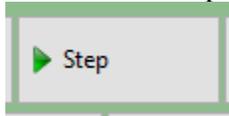
- d. Leave all the remaining options in their default settings.

55. The test mode in CellProfiler lets you see the image analysis happening one step at a time so you can observe how the image processing is working.

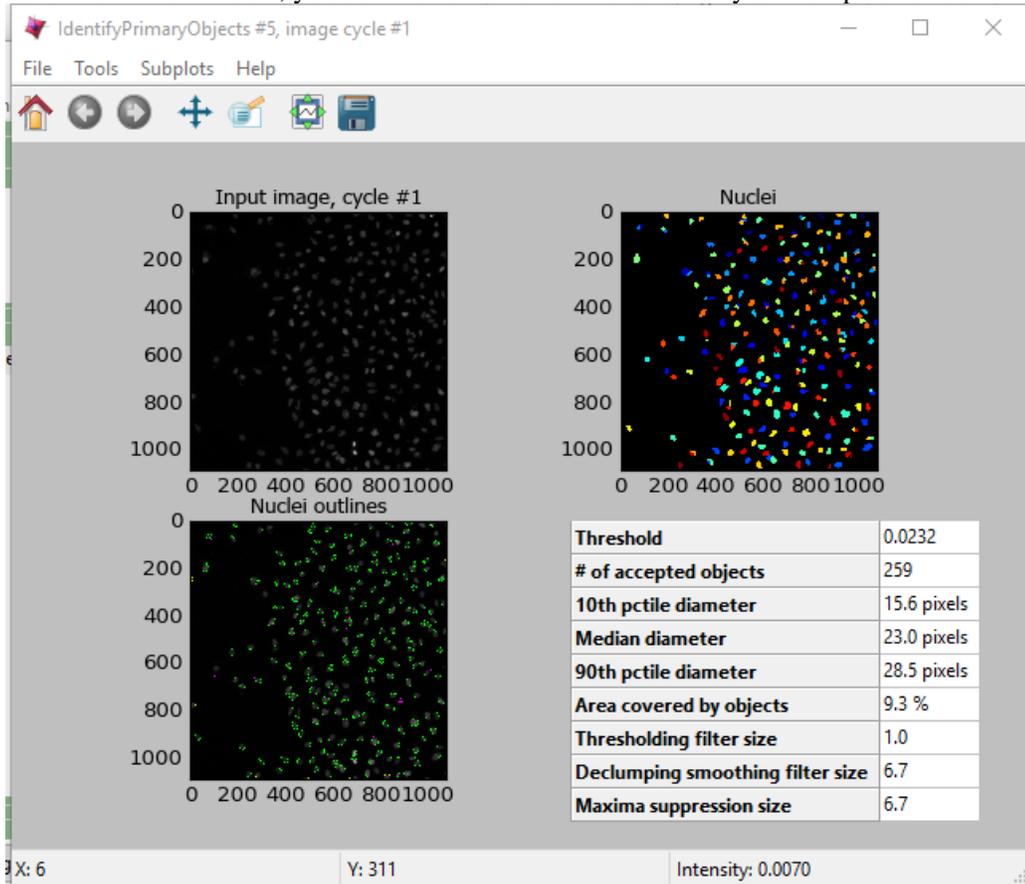
- a. Click on the “Start Test Mode” button in the bottom left window.



- b. Click on the “Step” button.



- c. You will see a new window appear that shows the input image, the primary objects identified as nuclei, the outlines of the objects, and summary data on the objects for that image. If you can't view this window, you can move it to a new location on your computer screen.



56. In this particular image, it appears that CellProfiler is accurately recognizing the nuclei. If it were not recognizing the nuclei, you could go back to the IdentifyPrimaryObjects module to adjust the parameters as needed.

Q12. Discuss the distribution of nuclei across the field of view. Are the nuclei randomly distributed? Are the nuclei separated from each other, or are any overlapping or touching? Do the nuclei all have the same shape and brightness?

57. Click the “Exit Test Mode” button.

#### 4.3.5 Set Up CellProfiler to Recognize and Segment Cells

58. You are now ready to tell CellProfiler how to recognize the cells. You will use the AGP images for this, since this label tends to identify many components of the cell:
- Go back to the NamesAndTypes module and click the “Add another image” button.
  - Set the “Select rule criteria Match” to select files containing “AGP” in the match criteria text box, and then enter “AGP” as the name to assign the image files.
  - Click “Update”. You should now see two columns: one for DNA and one for AGP.

59. You will now provide CellProfiler with instructions for how to recognize the cells in the control image data set, using the nuclei to help segment the cells:

- In the menu bar, select “Edit”, then “Add Module >”, then “Object Processing >”, then “IdentifySecondaryObjects”.
- This new module should appear in the pipeline after IdentifyPrimaryObjects. If it is not second, then either click and drag it down, or select it and click the down arrow button in the Adjust modules tools at the bottom of the window.



- In the Module notes window on the right, find the “Select the input image” field and select “AGP” from the drop-down menu.
- Find the “Select the input objects” field and select “Nuclei” from the drop-down menu. This tells CellProfiler to use the nuclei (identified in the prior module) to help identify and segment the cells from each other.
- In the “Name the objects to be identified” field, enter “Cells” (it may already be there as the default name).

60. Use the test mode again to view the pipeline one step at a time:

- Click on the “Start Test Mode” button.
- Click on “Step”. You should see the same IdentifyPrimaryObjects output window as last time.
- Click on “Step” again. You should now see the IdentifySecondaryObjects output window.
- Arrange the two output windows side-by-side to examine the results of CellProfiler using an algorithm to segment the AGP image into separate cells. Recall that in the previous exercise you intuitively developed your own rules to segment the cells. A summary of the CellProfiler process is shown below (Figure 3).

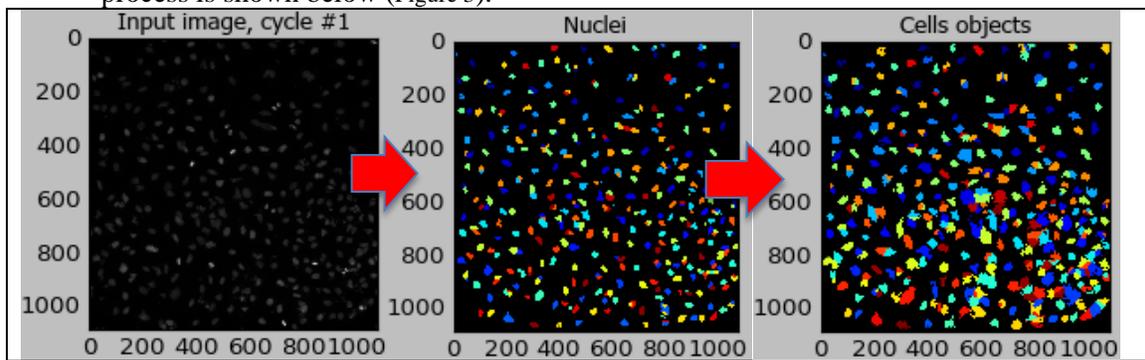


Figure 3: One original uploaded image and its corresponding nuclei and cells as identified by CellProfiler

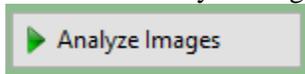
61. If you’d like to use the test mode to cycle through another image set, click the “Next Image Set” button and then the “Step” button. When you are done looking at additional image sets, click on the “Exit Test Mode” button.

#### 4.3.6 Set Up CellProfiler to Measure and Export Cellular Phenotypes

62. Now that CellProfiler has identified the nuclei and cells, you can instruct the software to count the cells and quantitatively determine various morphological characteristics of the nuclei and cells, which represent cellular phenotypes:

- In the menu bar, select “Edit”, then “Add Module >”, then “Measurement >”, then “MeasureObjectSizeShape”.
- Move this new module to the end of the pipeline, if it is not already last.

- c. In the “Module notes” window for this new module, in the “Select objects to measure” drop-down, select “Nuclei” from the drop-down menu.
  - d. Click the “Add another object” button and then select “Cells” as the objects.
  - e. In the “Calculate the Zernike features?” option, select No. You won’t need these calculations today, so this will reduce the processing required by your computer, which will speed up the analysis.
63. Now you need to instruct CellProfiler to output the results in a format that you can use in other applications.
- a. In the menu bar, select “Edit”, then “Add Module >”, then “File Processing >”, then “ExportToSpreadsheet”.
  - b. Move this new module to the end of the pipeline, if it is not already last.
  - c. In the “Module notes” window for this new module, go to the “Filename prefix” field and enter “Control\_”.
  - d. In the “Select the measurements to export” option, select “Yes”, and then click on the “Press button to select measurements” button that will appear. This will open the “Select measurements” window. In this window, click on the following:
    - i. Under Cells > AreaShape, check Area.
    - ii. Under Image > Count, check Cells.
    - iii. Under Nuclei > AreaShape, check Area.
    - iv. Click on the “OK” button
  - e. In the “Calculate the per-image mean values for object measurements?” option, select “Yes”. This will tell CellProfiler to export the mean values across cells in each image, which is what you will use for the statistical analysis.
64. You’re now ready to run an automated analysis of all of your control images:
- a. Select the “Analyze Images” button (instead of the Start Test Mode” button).



- b. CellProfiler will open four output windows, which will each be updated as the analysis proceeds. Note that CellProfiler indicates the elapsed processing time, an estimate of the total processing time in the lower, and a progress bar in the lower-right corner of the main window.
- c. CellProfiler will display a message when the processing is complete. Click OK to close this message.

### 4.3.7 Analysis of Mutant Allele Phenotypes

Now that you have acquired cell count, cell area, and nuclear area data on the control cells, you will perform the same analysis on the images of cells with the mutant alleles.

65. Add the Allele 1 images to the CellProfiler file list:
  - a. Click inside the “File list” window in the main CellProfiler window.
  - b. In the menu bar, select “Edit”, then “Select All”. This should highlight all of the image files in the list.
  - c. Press the “Delete” key on your computer keyboard to remove all the files from the list. (This only removes the files from CellProfiler, it does not remove them from your computer.)
  - d. You should now see the “Drop files and folder here” message. If you do not, you still have undeleted files in the list.
  - e. Locate the EGFR\_Mutant\_1 folder in the Mutation\_Screen\_Images source folder and drag it to the File list window in CellProfiler.
66. Go to the Module notes window for the NamesAndTypes input module and click the Update button. You should see the AGP and DNA images populate the two-column table.

67. Go to the Module notes window for the ExportSpreadsheet analysis module and change the filename prefix to “EGFR\_Mutant\_1\_”.
68. Click on the “Analyze Images” button.

#### 4.4 STATISTICAL ANALYSIS

You will now determine whether the mutation in the EGFR gene significantly affected the nuclei number or nuclear area. To do this, you will use a t-test to calculate the probability that any difference in the mean number or mean area between the control and mutant treatments is due to chance.

- Q13. For the purposes of statistical analysis, can each of the nine sites (images) within each replicate be considered a replicate? That is, if you determined the mean nuclei count for nine images for each of five replicates, do you have 45 replicate measurements or only five replicate measurements? Explain your answer.

##### 4.4.1 Determine the Replicate Means

Your output data are in two relevant files. To analyze the data in a spreadsheet application, you must first reformat your data into a single file with each row representing an image (site), and with seven columns: Replicate, Control Cell Count, Control Cell Area, Control Nuclei Area, Mutant Cell Count, Mutant Cell Area, Mutant Nuclei Area.

69. Go to your “Analysis Output” folder and open your two image data output files (these should be named Control\_Image.csv; and EGFR\_Mutant\_1\_Image.csv).
70. Create a new worksheet and save it in the Analysis Output folder as “Mutation\_Screen\_Summary.xls”.
71. In the Mutation\_Screen\_Summary file, starting with cell A1, input the following in the first row as column titles: Replicate, Control Cell Count, Control Cell Area, Control Nuclei Area, Mutant Cell Count, Mutant Cell Area, and Mutant Nuclei Area.
72. Copy and paste the relevant data for cell count, cell area, and nuclei area from the Control\_Image file to appropriate columns in the Mutation\_Screen\_Summary file.
73. Copy and paste the relevant data for cell count, cell area, and nuclei area from the EGFR\_Mutant\_1 file to the appropriate columns in the Mutation\_Screen\_Summary file.
74. Save the Mutation\_Screen\_Summary file.

##### 4.4.2 Test for Differences Between Means

75. Perform a two-tailed t-test for count data in the Data Analysis spreadsheet. If you are unsure how to do this within the spreadsheet application, review [Spreadsheet Tutorial 4](#) (t-test) on the HHMI BioInteractive [Spreadsheet Data Analysis Tutorials](#) site.
- Q14. Did you observe a significant effect of the EGFR mutation on the mean nuclei count? Did the result match your prediction?
76. Repeat the preceding steps for the nuclear area data.
- Q15. Did you observe a significant effect of the EGFR mutation on the mean nuclei area? Did the result match your prediction?
77. When you are done, save your Data Analysis spreadsheet.

## 5 INSTRUCTOR NOTES

Sample analysis data and all required fluorescence images are available here:

<https://www.dropbox.com/sh/umwth5uwtbv2zg1/AADrDYFQJvk82VVeUh0LLOfya>.

This activity uses images in the [Mutation\\_Screen\\_Small\\_Set](#) folder. Students can explore additional images of the same cell line with different oncogene mutations in the [Mutation\\_Screen\\_Full\\_Set](#) folder.

It is recommended that the instructor download the desired image set folder and copy this folder to each of the student computers before the students begin the activity.