Reference Search, Image and Data Analysis

Artem Oganesyan and Narine Sarvazyan

Contents

N. Sarvazyan (ed.), *Tissue Engineering*, Learning Materials in Biosciences,
https://doi.org/10.1007/978-3-030-39698-5_1

What You will Learn in This Chapter and Associated Exercises

Students will become familiar with most common databases and Internet search engines that can help them to find and collect information required in this course. They will also learn how to perform basic data analysis and extract quantitative information from acquired images of cells and tissues.

1.1 Literature Search

In today's world where a great amount of data can be digitally stored and readily retrieved, literature search becomes one of the most essential skills that every researcher should own. Below we briefly overview the main search engines that can be used to extract relevant information needed by students to complete their assignments throughout the course.

PubMed is probably one of the best-known free search engines when it comes to biomedical literature. Launched in June of 1997, PubMed gives free access to the Medical Literature Analysis and Retrieval System (MEDLINE)—a large bibliographic database of references, abstracts, and, in many cases, full-text articles on life sciences and various biomedical topics. This database is maintained by the United States National Library of Medicine. As of winter of 2020, PubMed contained over 30 million entries, with half a million new records added each year. One of the biggest assets of PubMed is its search tool. On the left side of the page, users can manage their searches by typing in the category of article, the date of publication, and many other variables. These additional filters, such as language, age of study participants, animal versus human studies, or journal categories, are aimed to narrow down the search. It can be then further improved with sorting articles by "Best Match," "Most Recent," or "First Author" just below the search bar. On the right side of the chosen article page, the user can also see relevant content under "Similar Articles" heading. Once users find the articles that they were looking for, they can send them to a virtual clipboard or store them in "My Bibliography." By using a reference management software (discussed below), the entire content of the user's clipboard can be then added at once to their reference library.

Google Scholar is another widely used free web search engine. Compared with PubMed, it focuses on an even larger array of disciplines, including non-medical sciences (e.g., social sciences, economics, computer science, arts). In addition to this, besides peer-reviewed journal articles, larger literature sources are included in the Google Scholar engine, such as numerous textbooks, conference reports, dissertations, patents, and viewpoints. For beginners, Google Scholar might seem more user-friendly than PubMed since its search tool resembles one of Google's. Nevertheless, searches in Google Scholar tend to be less specific because of the absence of different search strategies and tools provided by PubMed. On the other hand, Google Scholar has a very useful feature (currently absent in PubMed) that shows how many times a particular article was cited and by whom. By clicking on articles that cite the one user just read, one is able to trace later publications on the same subject.

Another rich source of scientific information, including detailed TERM protocols, can be *published patents and patent applications*. They can be found on ► USPTO.com website or the more recently developed online platform ► www.lens.org. Both provide free access to millions of worldwide patents and are fully searchable. The ► Lens.org website also includes access to scholarly articles and enables to use graphics to display results of the searches in a more effective visual way.

1.2 Reference Management Software

Keeping track of multiple publications is a cumbersome task. In order to address this need, several commercial software applications have been developed, common examples being *EndNote* or *Reference Manager*. These reference management programs helped users to collect, store, and manage large numbers of research papers. Yet, these programs were quite expensive, so a group of proactive graduate students from Germany developed a free alternative they called *Mendeley* (after the biologist Gregor Mendel and chemist Dmitri Mendeleyev). Their platform became so popular that in 2013 Mendeley was acquired by Elsevier for over 50 million dollars. One of the sale conditions was that Mendeley continues to be free to users unless a large amount of PDF files is stored with it (to avoid this user can simply unclick "Save PDF" pop-up). One of the most useful functions of Mendeley and other reference management programs is their ability to create a Bibliography (also called Reference list) according to the guidelines of specific journals or publishers. When added to the browser as a Mendeley extension, online papers can be easily imported into a user's reference library to be later inserted in a Word document. Today, Mendeley can run on multiple platforms, including Windows, Mac, iOS, Linux, and Android.

1.3 Literature Access

Most publishing houses do not provide immediate free access to the scientific articles they publish. This is considered by many researchers to be an unfair practice detrimental to the progress of humankind yet very lucrative for scientific publishers. In 2014, the National Institutes of Health (the NIH—an organization that funds most of the biomedical research in the United States for a total annual cost of over 30 billion dollars) required all publishers to release, 12 months after their publication date, the content of the articles that were funded by the NIH. This was a very helpful step, yet it still left a large number of articles unavailable for researchers to read without paying high fees. This list included the most recent ones, the ones that were published before NIH implemented their 12-month policy, and the ones that were not directly funded by the NIH. Overall, out of 14.2 million articles on PubMed that have links to their full-text, only 3.8 million articles are free for any user.

To fight the high cost of article access established by publishing houses, a graduate student from Kazakhstan, Alexandra Elbakyan, started *Sci-Hub*. This website enables free access to scientific articles uploaded there by the users, and as of January 2020, over 78 million papers have been deposited to Sci-Hub [1]. By entering the title, URL, PMID, or DOI of the desired paper in the search bar on the website page, one can download articles in the PDF format bypassing publisher's paywalls. Sci-Hub domains have been periodically suspended due to legal battles with major publishers, and the website address is changing time after time. Yet, by now, Sci-Hub has become a critical working tool for many researchers, especially those in the developing world. In 2016, Alexandra Elbakyan was named by *Nature* magazine to be one of the ten most influential people in science.

Today there is a growing global movement to make science advances available to every person on the planet regardless of their financial status. In August of 2018, the European Research Council decided to make all European scientific publications freely accessible by the year 2021, requiring publishers to remove paywalls and other sorts of obstacles to download journal content (www.coalition-s.org). One can hope that similar trends will be followed by other governmental bodies ushering a new era of scientific information exchange.

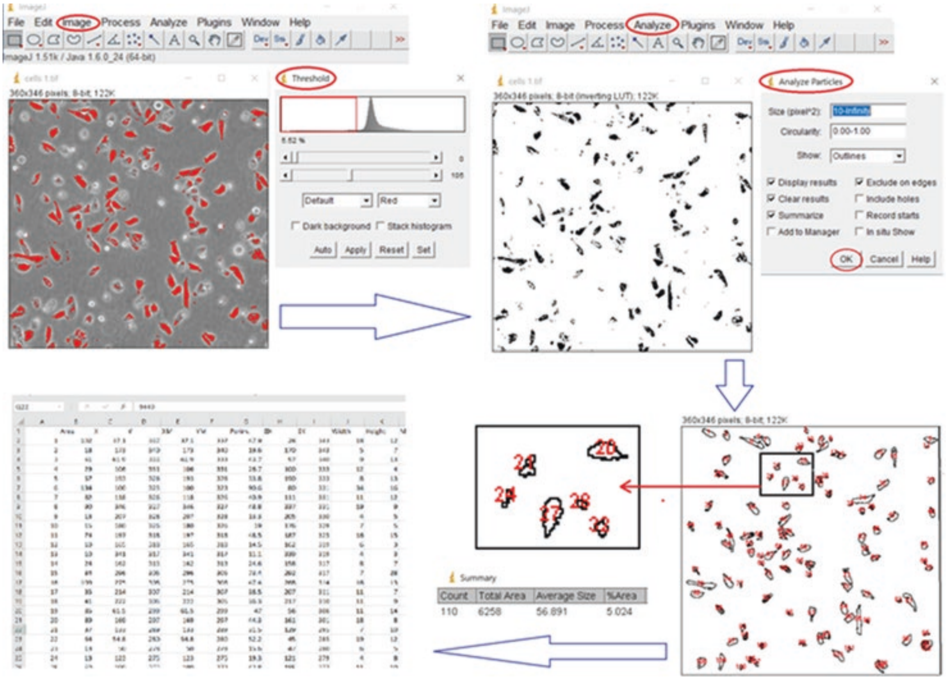
1.4 Video Journals

The most convenient way to learn new protocols is to observe them. This can be done remotely by using Youtube, Protocol Exchange, and other searchable video sites. There are also several new journals that publish experimental methods in video format. The most known is called JoVE (*Journal of Visualized Experiments*). JoVE was founded by a frustrated graduate student who tried multiple times to repeat her predecessor's experiments without much luck, and so to address her frustrations she created this new venue. JoVE publishes videos of research protocols from many fields, including tissue engineering. Unfortunately, due to JoVE's overwhelming success, users' access to the videos these days is not always free as JoVE turned into a commercial powerhouse. However, many of the earlier video protocols are still freely available and can be very useful to complete the team's assignments.

1.5 Image Analysis

Statistically valid image analysis is one of the key ingredients to derive proper results and conclusions. Various software programs can help to facilitate this process. One of the most useful and widely used programs was developed by the United States National Institutes of Health (NIH) and is called *ImageJ* [2]. It is free to users and is available on different operating systems, such as Microsoft Windows, MacOS, and Linux. ImageJ has numerous features, and students are encouraged to explore them on their own. There are also multiple online ImageJ tutorials, videos, and articles that explain how a particular parameter can be measured from a digital image. The software allows users to work with different types of images and enables easy editing, analyzing, processing, and storing 8-bit color and grayscale, 16-bit integer, and 32-bit floating-point pictures. ImageJ can read a myriad of various image file formats, including common ones such as PNG, GIF, JPEG, or TIFF or more rare types such as LSM. This can be of particular importance as images obtained from different microscopes and other equipment can have different file extensions.

ImageJ program can help users to measure distances and angles in any given picture or to calculate a specific area within the image. Users can choose certain thresholds for the intensity of selected objects, based on which the program will then collect statistics of different objects' values. After thresholding and object selection, ImageJ can also sort different objects depending on their size, shape, perimeter, or signal intensity (◘ Fig. 1.1). Common geometric affine transformation (e.g., rotation, flips, translation, or scaling) is also possible using ImageJ. The program allows the processing of arithmetical operations between different images. Another useful feature of ImageJ software is its ability to produce histograms, line charts, or 3D rotatable graphs for visual data presentation.


ImageJ
File Edit Image Process Analyze Plugins Window Help
ImageJ 1.51k / Java 1.6.0_24 (64-bit)
cells 1.tif
360x346 pixels; 8-bit; 122K
Threshold
5.62 %
0
106
Default
Red
Auto
Apply
Reset
Set
360x346 pixels; 8-bit (inverting LUT); 122K
Analyze Particles
OK
Cancel
Help
Summary
360x346 pixels; 8-bit; 122K
28
26
29
28
27
30

□ Dark background □ Stack histogram

Size (pixel^2): 10-Infinity
Circularity: 0.00-1.00
Show: Outlines

☑ Display results ☑ Exclude on edges
☑ Clear results □ Include holes
☑ Summarize □ Record starts
□ Add to Manager □ In situ Show

Count	Total Area	Average Size	%Area
110	6258	56.891	5.024

Fig. 1.1 An example of an original and thresholded image of plated cells and a range of different parameters that can be extracted by the ImageJ program from each thresholded object

$$\mu_x = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$SD = \frac{\sqrt{\sum_{i=1}^{n} (x_i - \mu_x)^2}}{n-1}$$

$$SEM = \frac{SD}{\sqrt{n}}$$

Fig. 1.2 Formulas behind basic statistical terms

1.6 Statistical Analysis

Another set of basic skills that students will be required to apply in this course includes a statistical analysis of a given data. Students are referred to various online resources to refresh their memory of basic statistical terms. Here we will just briefly mention the few most commonly used terms (Fig. 1.2).

Variance and Standard Deviation (SD) Variance is a principal measure that shows the variability of a quantity. It is calculated by squaring the deviations from the mean and then averaging these squares. Squaring enables one to include not only positive deviations but also negative ones. By taking the square root from the variance, users can calculate the standard deviation (SD), which has the same units as the initial data.

The *Range of a Variable* is the difference between the largest and the smallest given data. Keep in mind that the value of the range depends on the number of experimental points. In other words, the bigger the number of observations, the larger the range of that variable is likely to be.

The *Standard Error of the Mean* (SEM) shows how far the mean derived from *N* number of samples is from the mean of the entire population. To calculate SEM, first the user calculates the standard deviation as per the formula above and then divides it by the square root of *N*.

1.7 Understanding the Key Difference Between SD and SEM

It is important for readers to understand the main difference between SD and SEM. Let us say we have a high school consisting of 1200 students. Ten students decided to know their height in absolute numbers. These measured numbers (in centimeters) ended up being: 168; 161; 183; 172; 193; 185; 170; 165; 175; 158. The easiest value to obtain from this sequence is the range of the variable—the difference between the highest and lowest values of our sample, which is 193–158 = 35. In order to know the mean of the sample, we need to sum all these values and divide it by $N = 10$ (called sample size). In our case, the mean will be 173. Once we get the mean, we can calculate SD. The sum of squared deviations from the mean will be 1116. To find SD we will have to divide that number by the sample size minus 1, followed by taking square root from that number. This operation will yield SD = √(1116/(10–1)) = 11.13. In order to implement identical SD calculations in the Microsoft Excel program, students can use the following formula SD = stdev(data range).

Now, let's say we want to know how far is that sample mean from the mean height of *all* the students in the entire school. The standard error of the mean is then calculated by dividing SD by the square root of $N = 10$ yielding SEM = 3.52. To implement SEM calculation in Excel, one can use the following formula SEM = (stdev(data range))/SQRT(count(data range)).

Looking at the SEM formula (◘ Fig. 1.2), one may conclude that the SEM value will be smaller when more students are included in the calculation. In other words, the more samples that are drawn from the entire school population, the closer the mean will be to the actual mean height value. In our particular example, this value can be ultimately obtained by measuring the height of *all* 1200 students. It will just take more time. In real-life experiments though, measuring *all* the subjects, being them cells, animals, or human patients, is simply not an option. Therefore, by sampling more (i.e., doing more experiments) the user is able to get closer and closer to the ideal mean population value. SEM basically shows how far we are from it.

Note that SEM is directly proportional to the SD or variability of mean values between samples. This means that if the mean heights significantly vary between the students, more sampling (i.e., more *N*) is required to better estimate a true mean value of a population.

It is important for students to understand the fundamental differences in using these two measures. SD points to how wide are the differences between the individual measurements. SEM shows how close the mean value derived from the multiple measurements is to the mean of the whole population. More about this topic can be found in [3].

1.8 Implications of the Above Statistics for Students' Experimental Design

For future experiments in this course, students should draw conclusions based on at least three independent experiments, each done using at the minimum triplicate measurements. Triplicate means that the same measurement is repeated three times using cells or tissues from the same prep. For example, let's consider how to analyze a possible experiment designed to compare attachment of hepatocytes to laminin versus gelatin-coated dishes. First, a user needs to plate cells on at least six coverslips: three for laminin and three for gelatin group (this will be triplicate measurements for each coating type or $n = 3$). Then the same experiment has to be performed, let's say, four different times using different preps ($N = 4$). Analysis of samples using ImageJ tools enables the user to derive the percentage of cell coverage. The next step is to find average values from triplicates acquired during each of the four experiments (see ◘ Table 1.1). The next step will be to use the Microsoft Excel program (or its analog) to find mean and SEM for each type of coating. Excel t-test function can then be used to find out whether the two coatings were statistically different. In this case, it will be a paired t-test since each pair of mean measurements (mean values for laminin-coated dishes and mean values for gelatin-coated dishes) was acquired on the same day; therefore, they can be considered "paired." So, the final outcome of this experiment can be worded as: "Attachment of hepatocytes to laminin-coated coverslips was shown to be significantly higher compared to gelatin-coated coverslips. Specifically, expressed as mean±SEM, the percentage of surface cell coverage was 63.92 ± 5.61 vs 33 ± 3.48, $p < 0.05$, $N = 4$ with each individual experiment done in triplicate."

Students are encouraged to use online resources to understand the meaning of other key statistical parameters including correlation coefficient, ANOVA, and different forms of t-test. A good overview of basic statistics for presenting outcomes of the experiments can be found in this freely accessible reference [4].

◘ **Table 1.1** An example of an Excel sheet used to calculate the outcomes from four individual experiments ($N = 4$) each done in triplicate ($n = 3$)

Table	Treatment A				Treatment B				T-test
	#1	#2	#3	Avg	#1	#2	#3	Avg	*p*-value
EXP 1	50	45	24	39.67	67	46	58	57.00	
EXP 2	40	33	34	35.67	44	55	61	53.33	
EXP 3	20	47	33	33.33	81	76	78	78.33	
EXP 4	22	32	16	23.33	45	77	79	67.00	
			MEAN	33.00			MEAN	63.92	0.028
			SD	6.96			SD	11.21	
			SEM	3.48			SEM	5.61	

Session I

Demonstration

The instructor shows students the basic functions of PubMed, Sci-Hub, and Google Scholar as well as an example of how to use Mendeley and its Word plug-in to create a reference list. Then sample images are used to demonstrate key ImageJ functions and how to extract quantitative information and perform statistical analysis using basic Microsoft Excel functions.

Homework

Students are asked to download Mendeley and ImageJ software and familiarize themselves with their main features. Reading assignment: review TERM field history (Kaul & Ventikos, TISSUE ENGINEERING B, v.21, N.2, 2015 or its more recent equivalent).

Session II

Team Exercises

Each team selects a pair of sample images to be quantitatively compared. Any free online publication or other open sources can be used to obtain the two images. These can be cell cultures, decellularized tissues, scaffolds, or any other TERM relevant material. Teams then discuss visual differences between the two images and brainstorm how these differences can be translated into at least three quantitative values using ImageJ analysis tools.

Homework

Teams are tasked with extracting three different quantitative measurements from a selected pair of images using ImageJ tools. A short PowerPoint presentation is made by each team illustrating their quantification process and statistical analysis applied to the extracted data.

Sample Protocols

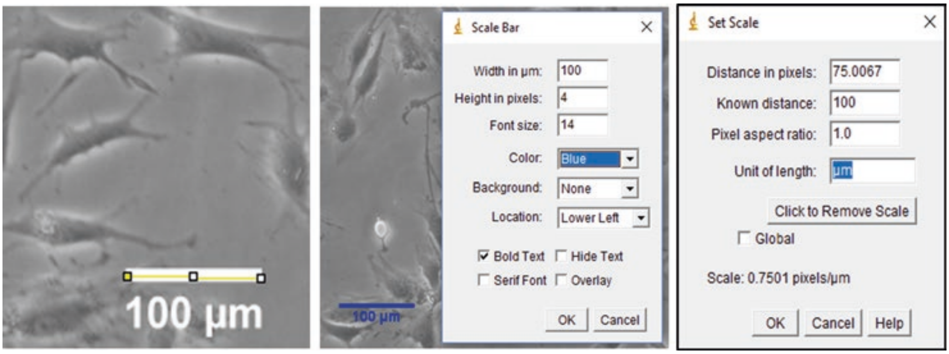
Below are examples of ImageJ steps that can be used to analyze images of cells or tissues.

Setting a scale for object measurements (◘ Fig. 1.3)

1. Trace the existing scale bar with the line selection tool.
2. Go to Analyze>Set Scale>Set known distance (e.g., 100) and distance in pixels (line length) and preferable unit (e.g., μm)>ok.
3. Go to Analyze>Tools>Scale Bar>Choose preferable features (width in μm, location, color, font size)>ok.

Counting objects

1. Select an area of interest by area selection tool (◘ Fig. 1.4).
2. Duplicate the image (Image>Duplicate).
3. Minimize the noise by subtracting the background (Process>Subtract the Background).
4. Convert the image into a black-white 8-bit type (Image>Type>8-bit).
5. Set the threshold manually until the cells are well distinguished (Image>Adjust>Threshold).
6. Make the image binary (Process>Binary>Make Binary).



Scale Bar

Width in µm: 100
Height in pixels: 4
Font size: 14
Color: Blue
Background: None
Location: Lower Left
☑ Bold Text ☐ Hide Text
☐ Serif Font ☐ Overlay

Set Scale

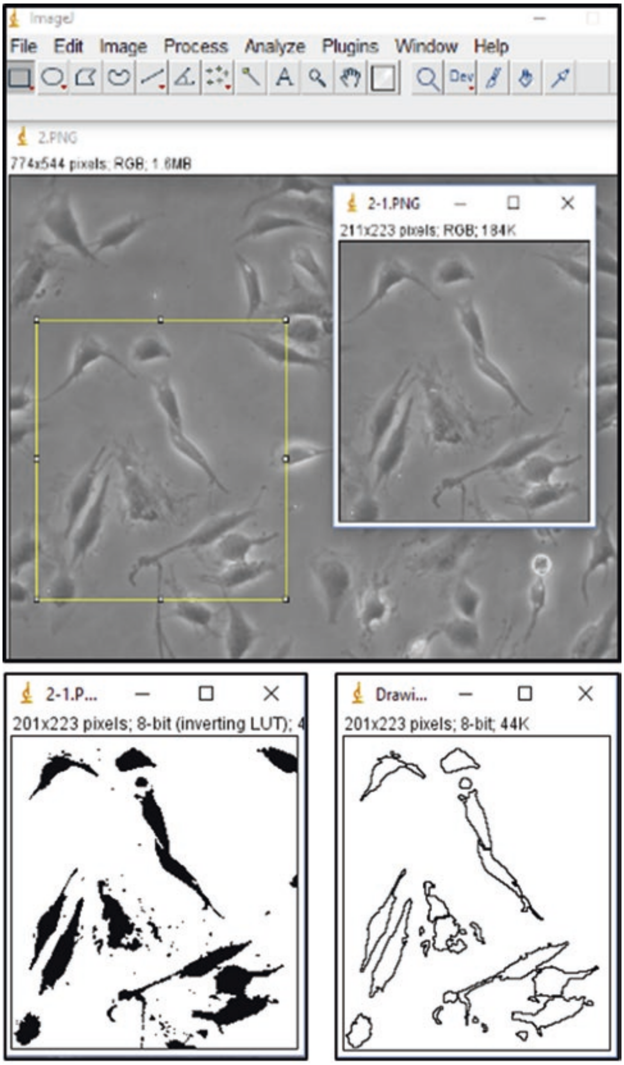
Distance in pixels: 75.0067
Known distance: 100
Pixel aspect ratio: 1.0
Unit of length: µm

☐ Global

Scale: 0.7501 pixels/µm

Fig. 1.3 Steps in ImageJ to calibrate an image for extracting quantitative measurements

Fig. 1.4 Use of ImageJ functions to outline and count cells



2.PNG
774x544 pixels; RGB; 1.6MB

2-1.PNG
211x223 pixels; RGB; 184K

2-1.P...
201x223 pixels; 8-bit (inverting LUT); 4

Drawi...
201x223 pixels; 8-bit; 44K

7. Fill holes (Process>Binary>Fill Holes).
8. Convert the image to mask (Process>Binary>Convert to Mask).
9. Separate confluent cells by watershed (Process>Binary>Watershed).
10. Count the cells (Analyze>Analyze Particles>Set a Minimum Particle Size (e.g., 10 µm)).

Area measurement (Fig. 1.5)

1. Calibrate the picture by setting a scale (see above).
2. Select the area of interest by the area selection tool.
3. Duplicate the image (Image>Duplicate).
4. Convert the image into a black-white 8-bit type (Image>Type>8-bit).
5. Adjust the threshold so that the desired area of measurement is black (Image>Adjust>Threshold).
6. Trace the outline of the area of treatment with the brush tool (Click on Wand (Tracing) Tool>Choose brush tool with right-click on oval/elliptical/brush tool).
7. Measure (Analyze>Measure).

This tool can be used to compare the timeline changes of the mean surface area of the same zone (i.e., on different days).

Area fraction

1. Change the type of image to 16-bit (Image>Adjust>16-bit).
2. Set an appropriate threshold (Image>Adjust>Threshold).
3. Set the required measurements: Area, Area fraction, and Display label (Analyze>Set Measurements).
4. Perform the measurement (Analyze>Measure).

The Area fraction tool can be used to compare, for example, the difference in collagen distribution between two sample images (Fig. 1.6).

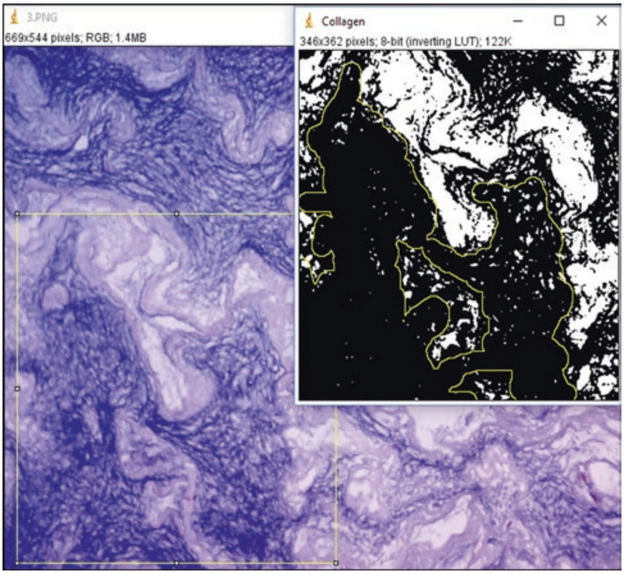
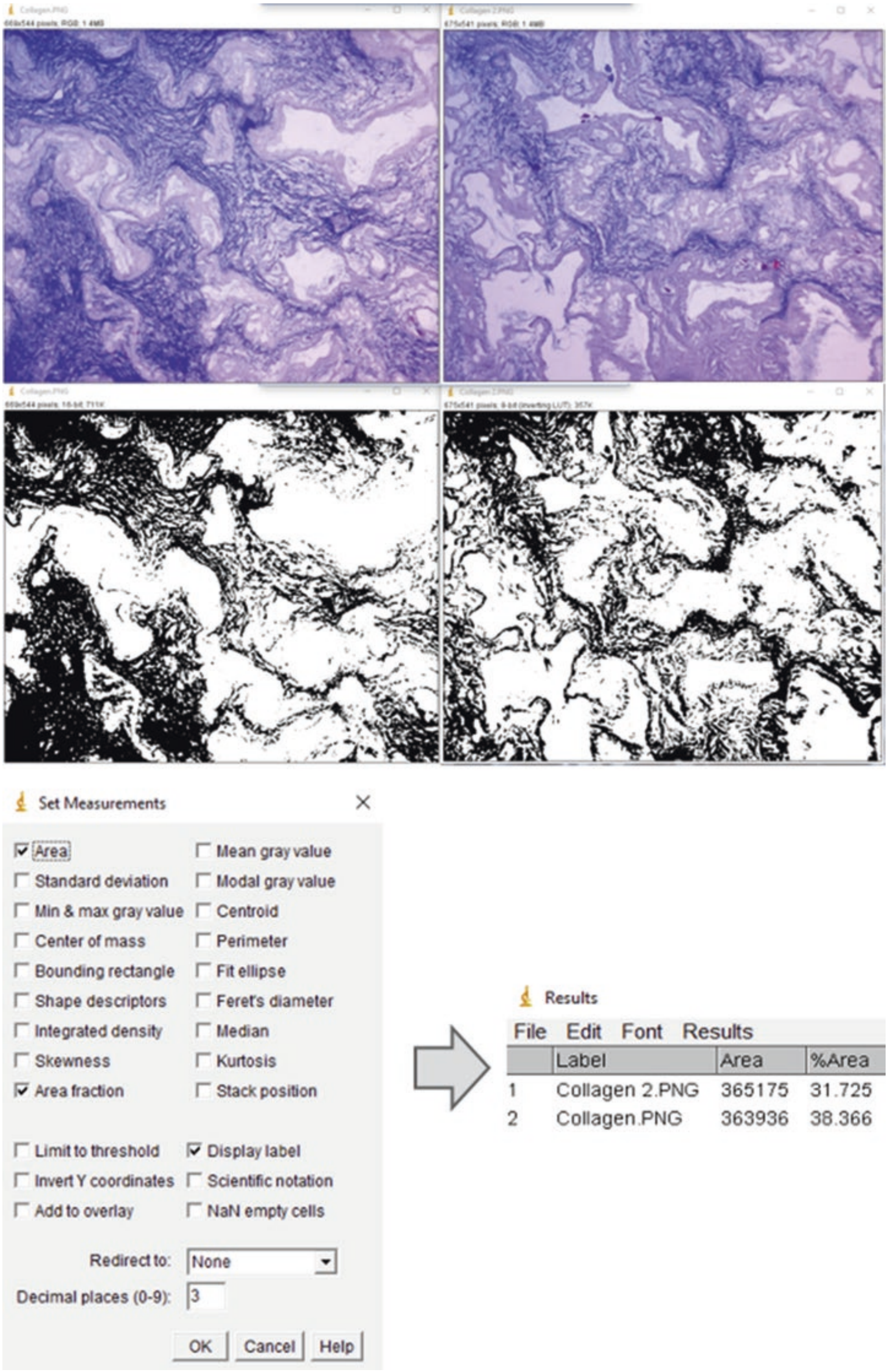

3.PNG
669x544 pixels; RGB; 1.4MB
Collagen
346x362 pixels; 8-bit (inverting LUT); 122K

Fig. 1.5 Use of ImageJ functions to manually outline and measure specific areas



Set Measurements

☑ Area
☐ Mean gray value
☐ Standard deviation
☐ Modal gray value
☐ Min & max gray value
☐ Centroid
☐ Center of mass
☐ Perimeter
☐ Bounding rectangle
☐ Fit ellipse
☐ Shape descriptors
☐ Feret's diameter
☐ Integrated density
☐ Median
☐ Skewness
☐ Kurtosis
☑ Area fraction
☐ Stack position

☐ Limit to threshold
☑ Display label
☐ Invert Y coordinates
☐ Scientific notation
☐ Add to overlay
☐ NaN empty cells

Redirect to: None

Decimal places (0-9): 3

Results

	Label	Area	%Area
1	Collagen 2.PNG	365175	31.725
2	Collagen.PNG	363936	38.366

Fig. 1.6 Screenshots illustrating ImageJ steps to quantitatively compare specific staining

Take-Home Message/Lessons Learned

After reading this chapter and performing the requested assignments and exercises, students should:

- Become familiar with general tools to search scientific literature including PubMed, Google Scholar, Lens.org, and other online databases
- Learn how to use reference management software, Mendeley being an example
- Become aware of availability of information contained in patents and patent applications
- Familiarize themselves with image processing software such as ImageJ and its basic functions
- Be able to use key statistical formulas to evaluate data and images quantitatively
- Understand the difference between standard deviation and standard error of the mean and its implication for study design

Self-Check Questions

? Q.1.1. You've read an interesting article and want to find papers that cited it since they can lead you to more recent work on the same subject. The best free online portal to find them is

A. Google Scholar
B. ImageJ
C. Sci-Hub
D. JoVE

? Q.1.2. You've read an interesting article and want to find similar papers on the same subject. The best free online portal to find them is

A. Google Scholar
B. PubMed
C. ImageJ
D. Sci-Hub

? Q.1.3. Free online portal with extensive visualization tools allowing to search for specific protocols within patent texts is

A. Google Scholar
B. ImageJ
C. Sci-Hub
D. Lens.org

? Q.1.4. Free software with multiple plug-ins enabling comprehensive image analysis is

A. Google Scholar
B. ImageJ
C. Sci-Hub
D. Lens.org

? Q.1.5. The standard error of the mean is __________ compared to the sample standard deviation.

A. A smaller value
B. A larger value
C. The same value
D. Impossible to tell

References and Further Reading

1. D.S. Himmelstein et al., Sci-Hub provides access to nearly all scholarly literature. elife **7** (2018). https://doi.org/10.7554/eLife.32822
2. C.A. Schneider, W.S. Rasband, K.W. Eliceiri, NIH image to ImageJ: 25 years of image analysis. Nat. Methods **9**(7), 671–675 (2012)
3. P. Barde, M. Barde, What to use to express the variability of data: Standard deviation or standard error of mean? Perspect. Clin. Res. **3**(3), 113 (2012)
4. G. Cumming, F. Fidler, D.L. Vaux, Error bars in experimental biology. J. Cell Biol. **177**(1), 7–11 (2007)