Semi-Automated Ground-Truth Data Collection and Annotation for Journal Figure Analysis

Rebekah Narum, Jie Zou, Sameer Antani
Communications Engineering Branch, National Library of Medicine, NIH, Bethesda, MD

NIH U.S. National Library of Medicine

OVERVIEW

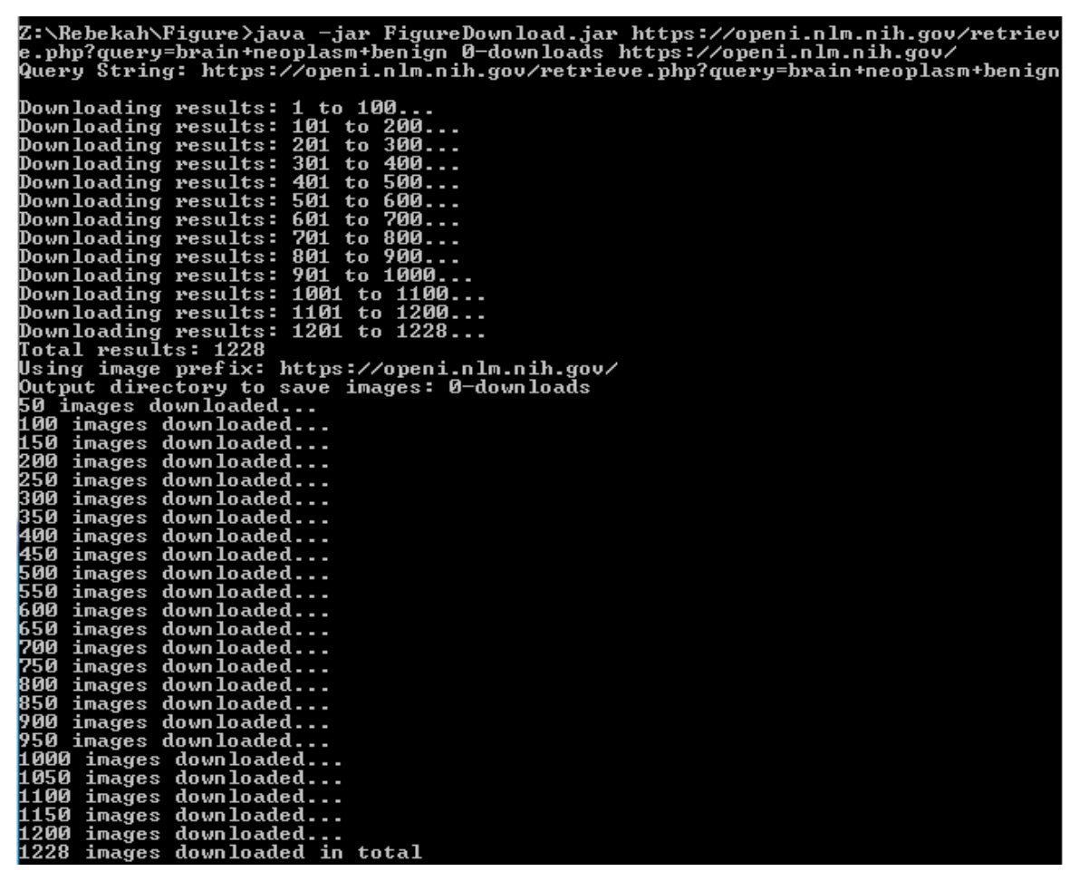
Open-i is an online service provided by the National Library of Medicine to enable search and retrieval of abstracts and images from 1.2 million PubMed Central® articles. An important preprocessing step of building Open-i backend is to automatically segment figures into panels and recognize panel labels, such that figure captions of individual panels can be linked to the figure panels and more precise features may be extracted. Existing panel segmentation and label recognition algorithms [1,2,3] are developed based on a tiny set of 448 figures. Due to the lack of training samples, the algorithms have to rely on many hand-crafted rules, which are not able to accommodate the large variations of the figures need to be processed.

This project creates a workflow pipeline, in an attempt to collect a significantly larger ground-truth annotated figure dataset efficiently. The annotation of a figure includes the style of the figure (single-panel, multi-panel or stitched multi-panel), rectangular bounding boxes of the panels, rectangular bounding boxes of the panel labels, and the panel labels. The work flow starts with running automated methods, and then the automated annotation is reviewed and fixed by humans. In order to ensure the annotation quality, a verification algorithm is developed to check the consistency of the annotations. The humans then review the suspicious annotations reported by the verification program.

FIGURE COLLECTION AND ANNOTATION

First, retrieve figures via Open-I Web Service API

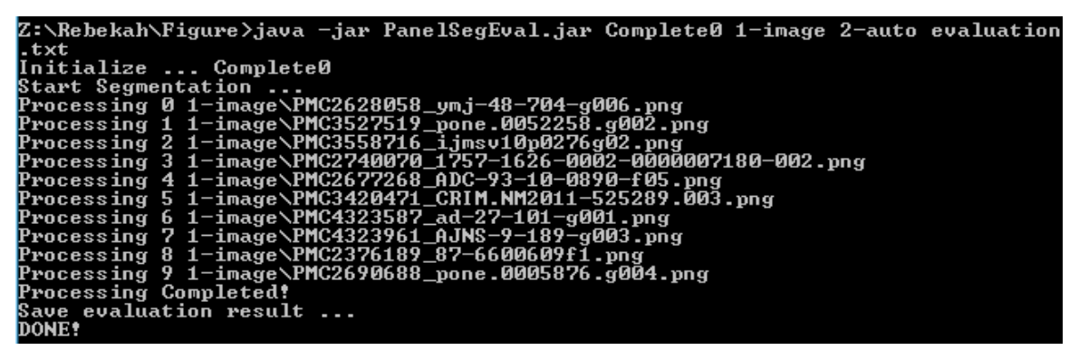
- 1000-2500 different images were downloaded at a time.
- The program FigureDownload (as shown in the image to the right) is called
  - the figures are automatically saved to the folder labeled 0-downloads.
- The query string used for downloading include:
  - turmeric,
  - heart + murmur,
  - broken + back,
  - brain + neoplasm + benign,
  - heart + block + congenital,
  - ovarian + tumor + benign,
  - ecg + brain,
  - ecg + heart + rate + low,
  - malaria + plasmodium + falciparum + with + complications.



```
Z:\Rebekah\Figure>java -jar FigureDownload.jar https://openi.nlm.nih.gov/retriev
e.php?query=brain+neoplasm+benign 0-downloads https://openi.nlm.nih.gov/
Query String: https://openi.nlm.nih.gov/retrieve.php?query=brain+neoplasm+benign

Downloading results: 1 to 100...
Downloading results: 101 to 200...
Downloading results: 201 to 300...
Downloading results: 301 to 400...
Downloading results: 401 to 500...
Downloading results: 501 to 600...
Downloading results: 601 to 700...
Downloading results: 701 to 800...
Downloading results: 801 to 900...
Downloading results: 901 to 1000...
Downloading results: 1001 to 1100...
Downloading results: 1101 to 1200...
Downloading results: 1201 to 1228...
Total results: 1228
Using image prefix: https://openi.nlm.nih.gov/
Output directory to save images: 0-downloads
50 images downloaded...
100 images downloaded...
150 images downloaded...
200 images downloaded...
250 images downloaded...
300 images downloaded...
350 images downloaded...
400 images downloaded...
450 images downloaded...
500 images downloaded...
550 images downloaded...
600 images downloaded...
650 images downloaded...
700 images downloaded...
750 images downloaded...
800 images downloaded...
850 images downloaded...
900 images downloaded...
950 images downloaded...
1000 images downloaded...
1050 images downloaded...
1100 images downloaded...
1150 images downloaded...
1200 images downloaded...
1228 images downloaded in total
```

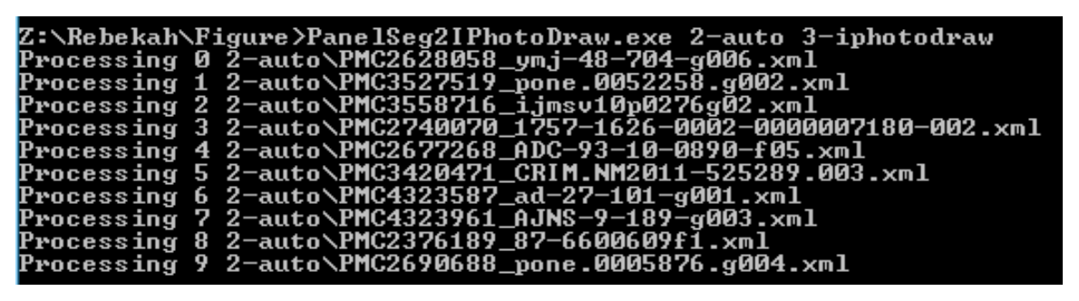
Second, generate automated annotation where the figures were manually copied from 0-downloads to a folder called 1-image, to limit the amount of images that are paneled and checked in one sitting.



```
Z:\Rebekah\Figure>java -jar PanelSegEval.jar Complete0 1-image 2-auto evaluation
.txt
Initialize ... Complete0
Start Segmentation ...
Processing 0 1-image\PMC2628058_ymj-48-704-g006.png
Processing 1 1-image\PMC3527519_pone.0052258.g002.png
Processing 2 1-image\PMC3558716_ijmsv10p0276g02.png
Processing 3 1-image\PMC2740070_1757-1626-0002-0000007180-002.png
Processing 4 1-image\PMC2677268_ADC-93-10-0890-f05.png
Processing 5 1-image\PMC3420471_CRIM.NM2011-525289.003.png
Processing 6 1-image\PMC4323587_ad-27-101-g001.png
Processing 7 1-image\PMC4323961_AJNS-9-189-g003.png
Processing 8 1-image\PMC2376189_87-6600609f1.png
Processing 9 1-image\PMC2690688_pone.0005876.g004.png
Processing Completed!
Save evaluation result ...
DONE!
```

Third, panel the images using an Automated Panel Segmentation tool named PanelSegEval.jar.

- The results are automatically saved to a folder labeled 2-auto.
- A small sample of the results of this step are shown on the right.



```
Z:\Rebekah\Figure>PanelSeg2IPhotoDraw.exe 2-auto 3-iphotodraw
Processing 0 2-auto\PMC2628058_ymj-48-704-g006.xml
Processing 1 2-auto\PMC3527519_pone.0052258.g002.xml
Processing 2 2-auto\PMC3558716_ijmsv10p0276g02.xml
Processing 3 2-auto\PMC2740070_1757-1626-0002-0000007180-002.xml
Processing 4 2-auto\PMC2677268_ADC-93-10-0890-f05.xml
Processing 5 2-auto\PMC3420471_CRIM.NM2011-525289.003.xml
Processing 6 2-auto\PMC4323587_ad-27-101-g001.xml
Processing 7 2-auto\PMC4323961_AJNS-9-189-g003.xml
Processing 8 2-auto\PMC2376189_87-6600609f1.xml
Processing 9 2-auto\PMC2690688_pone.0005876.g004.xml
```

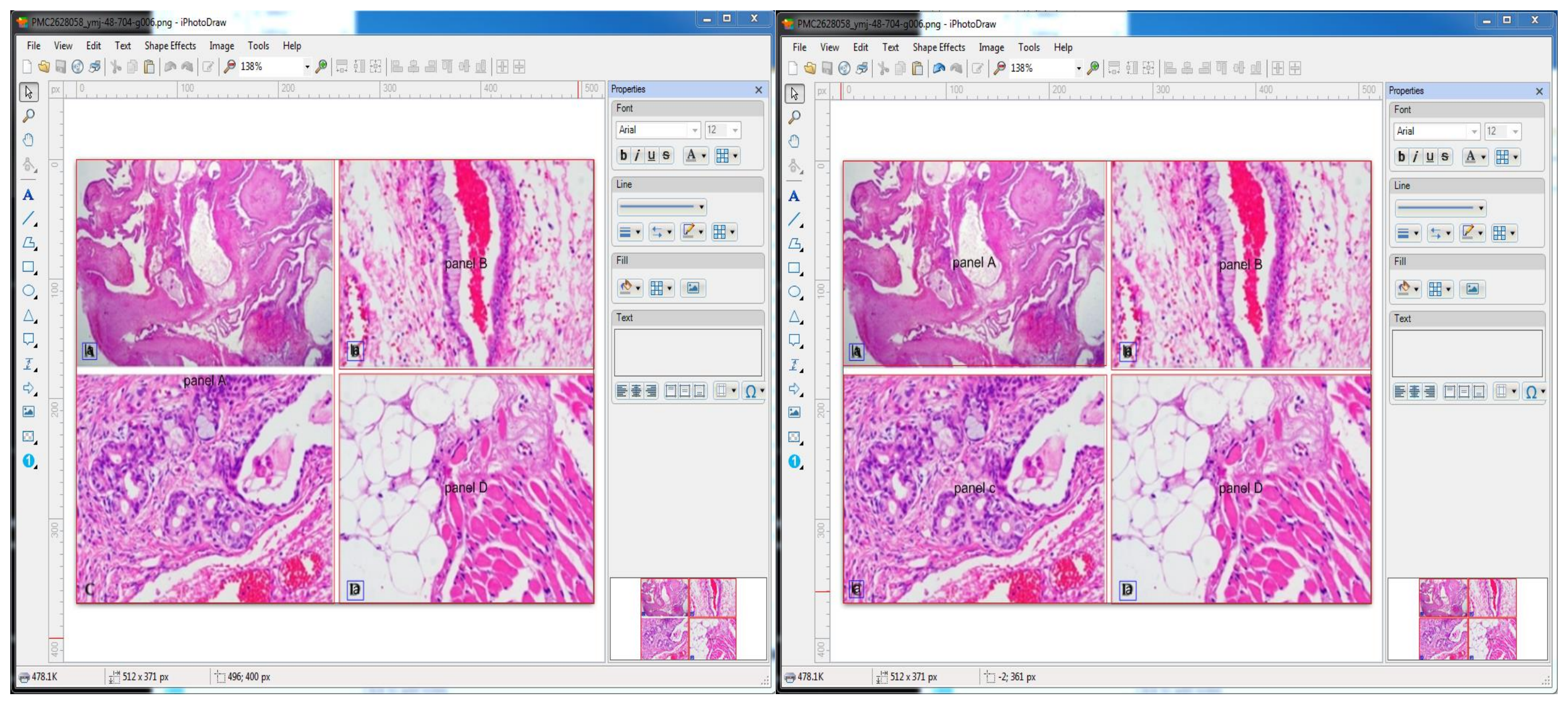
Fourth, the figures must be converted to an iphotodraw formatted XML file.

- The results are automatically saved to a folder labeled 3-iphotodraw.
- The XML files are copied and pasted into the 1-image folder.

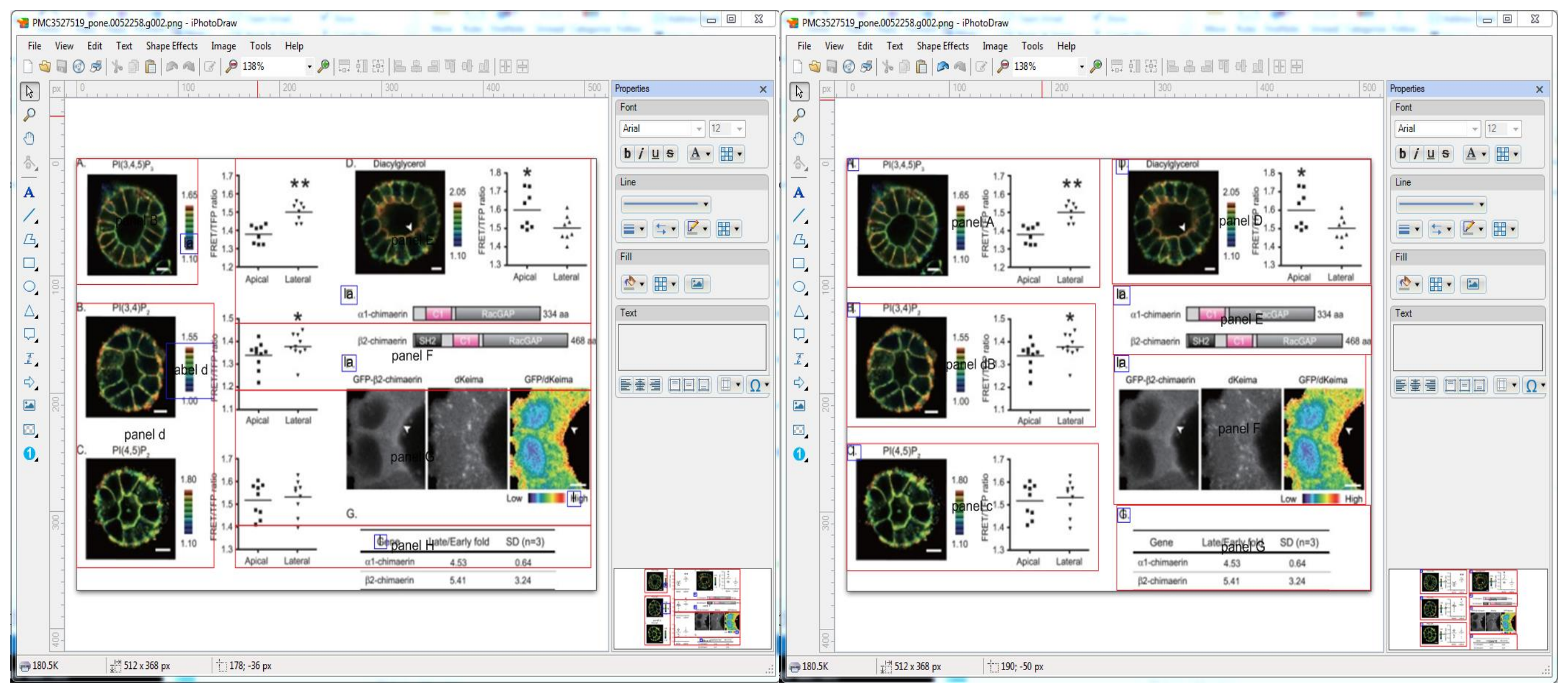
Fifth, use iPhotoDrawReconcile.exe in order to automatically open the figures, review, and correct the automated annotations made.

- iPhotoDraw is called, automatically displays the annotation(s), and allows for corrections until all images in the set are complete.
- The types of panels being annotated:
  - For a single panel figure, a panel bounding box marks the whole image.
  - For a multi-panel-without-label figure, the panel bounding boxes mark the individual figures.
  - For a multi-panel-with-label figure, the panel annotation is the same as a multi-panel-without-label figure except these panels must match up with their labels, e.g.: “panel A”, “panel a”, “panel B” should correspond with “label A”, “label a”, and “label B”.

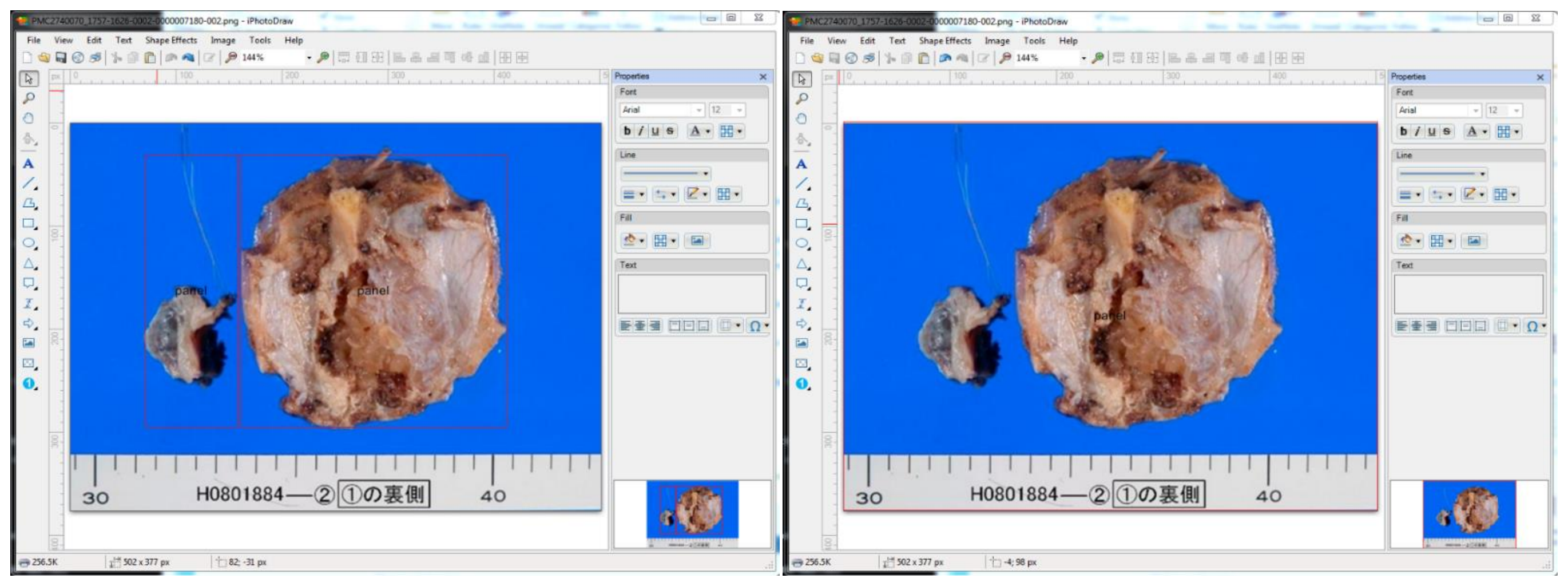
Finally, the figures are moved to a completed folder and the process starts over.



(A) This is a multi-panel image. The image on the left shows what the figure brings up automatically. The image on the right shows the figure after it has been corrected. This panel annotation example is relatively nice, only two errors, one being a missing panel box and the other being a missing label box.



(B) This is an example of a multi-panel image. The image on the left shows what was brought up automatically. The image on the right has been corrected. This is an example of a more typical figure because it has a lot more to fix.

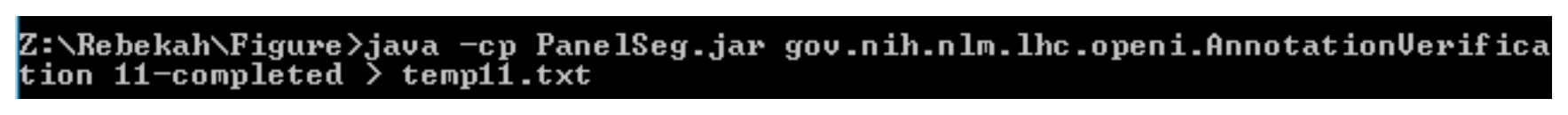


(C) This is an example of a single panel figure. The image on the left is the figure before being corrected. The edited figure is on the right.

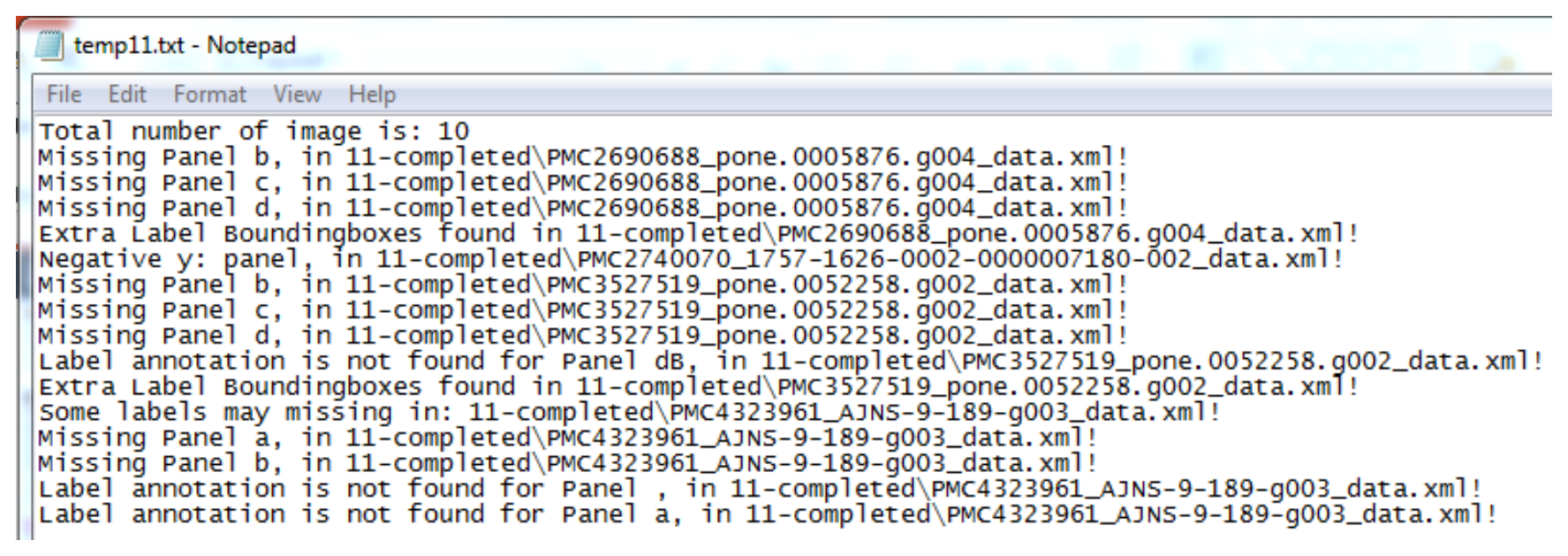
VERIFICATION

In this step, using the PanelSeg.jar program, the manual paneling corrections are checked for the following:

- That labels and panels match up
- That there are not any extra bounding boxes
- That there are not any missing labels or panels



```
Z:\Rebekah\Figure>java -cp PanelSeg.jar gov.nih.nlm.lhc.openi.AnnotationVerifica
tion 11-completed > temp11.txt
```



```
Total number of image is: 10
Missing Panel b, in 11-completed\PMC2690688_pone.0005876.g004_data.xml!
Missing Panel c, in 11-completed\PMC2690688_pone.0005876.g004_data.xml!
Missing Panel d, in 11-completed\PMC2690688_pone.0005876.g004_data.xml!
Extra Label Boundingboxes found in 11-completed\PMC2690688_pone.0005876.g004_data.xml!
Negative y: panel, in 11-completed\PMC2740070_1757-1626-0002-0000007180-002_data.xml!
Missing Panel b, in 11-completed\PMC3527519_pone.0052258.g002_data.xml!
Missing Panel c, in 11-completed\PMC3527519_pone.0052258.g002_data.xml!
Missing Panel d, in 11-completed\PMC3527519_pone.0052258.g002_data.xml!
Label annotation is not found for Panel dB, in 11-completed\PMC3527519_pone.0052258.g002_data.xml!
Extra Label Boundingboxes found in 11-completed\PMC3527519_pone.0052258.g002_data.xml!
Some labels may missing in: 11-completed\PMC4323961_AJNS-9-189-g003_data.xml!
Missing Panel a, in 11-completed\PMC4323961_AJNS-9-189-g003_data.xml!
Missing Panel b, in 11-completed\PMC4323961_AJNS-9-189-g003_data.xml!
Label annotation is not found for Panel , in 11-completed\PMC4323961_AJNS-9-189-g003_data.xml!
Label annotation is not found for Panel a, in 11-completed\PMC4323961_AJNS-9-189-g003_data.xml!
```

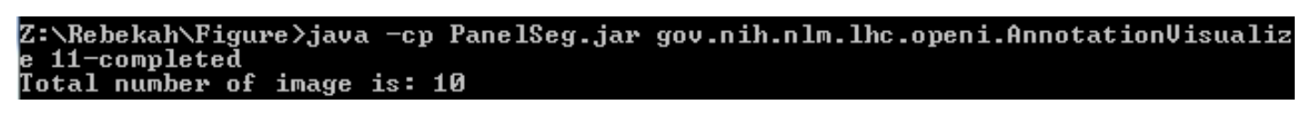
The individual figure names and the errors are saved to a txt file (as seen in the image to the right)

- The figure names are copied and pasted into iPhotoDraw for checking.

STYLE ASSIGNMENT

This is where the Style Annotation takes place.

- The program PanelSeg.jar is called to verify the annotations and determine the label styles.
- The label styles include the following:
  - Single-Panel is a single figure.
  - Multi-Panel has multiple figures in a given image which are clearly defined.
  - Stitched-Multi-Panel also has multiple figures in a given image, but the is no definitive line separating them.



```
Z:\Rebekah\Figure>java -cp PanelSeg.jar gov.nih.nlm.lhc.openi.AnnotationVisualiz
e 11-completed
Total number of image is: 10
```

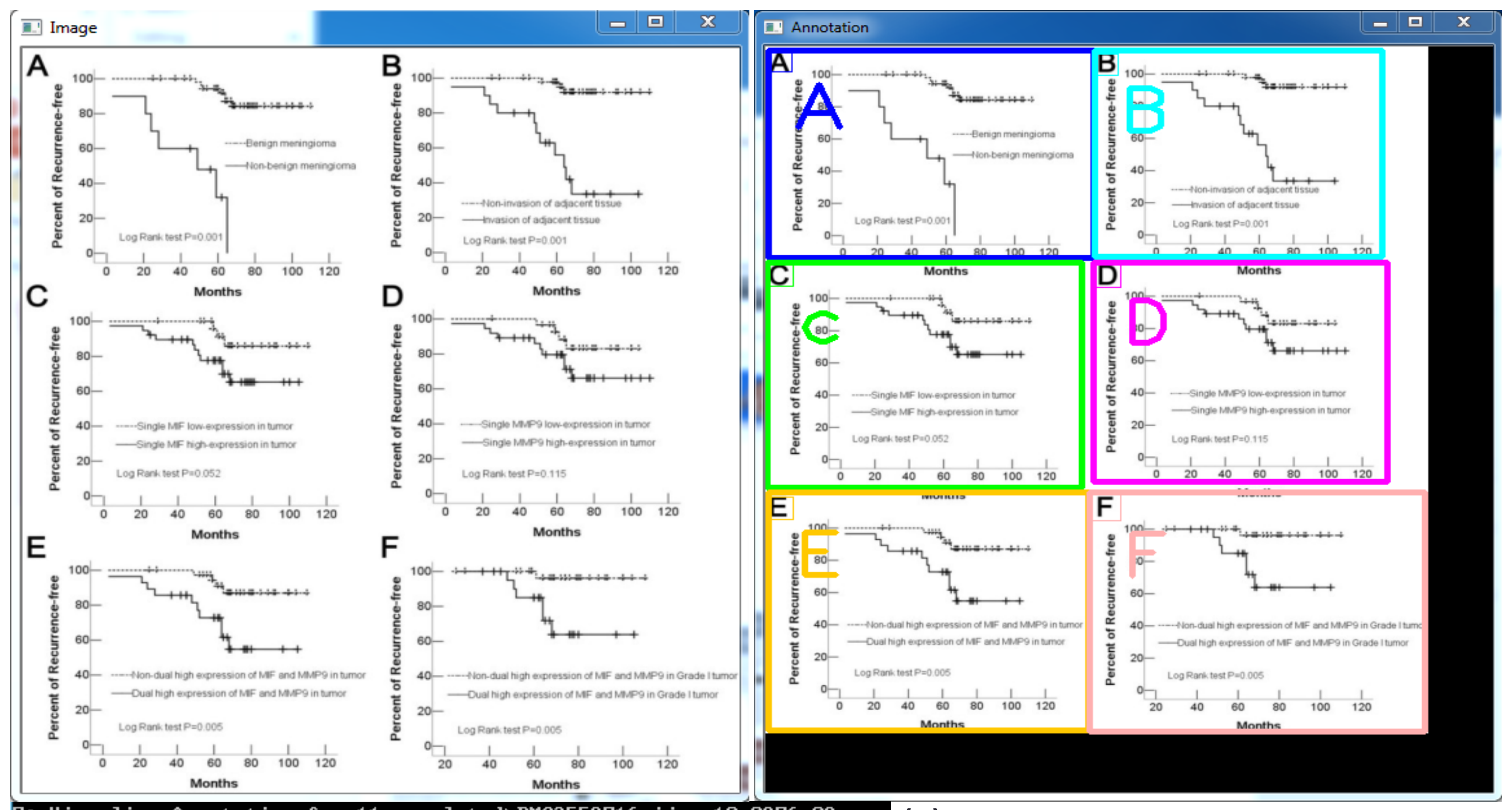
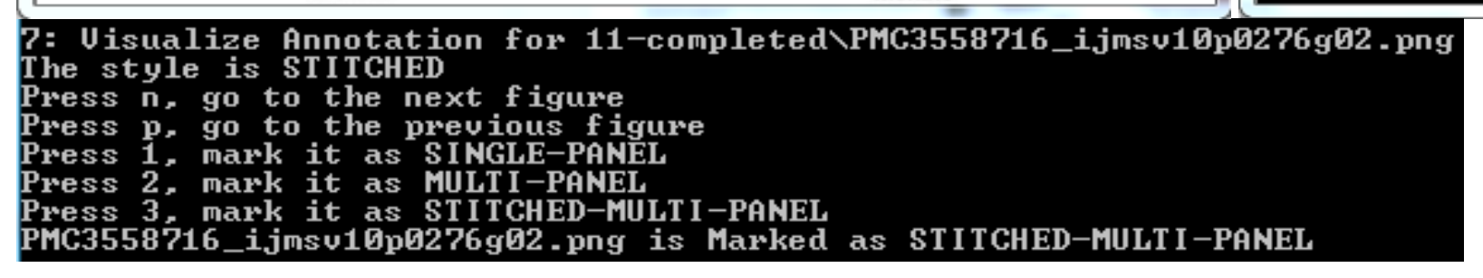
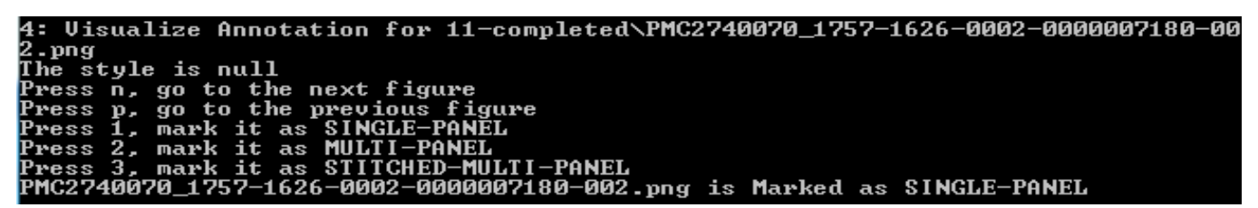


Image
Annotation
A
B
C
D
E
F
Percent of Recurrence-free
Months
Benign meningioma
Non-benign meningioma
Non-invasion of adjacent tissue
Invasion of adjacent tissue
Non-dual high expression of MIF and MMP9 in tumor
Dual high expression of MIF and MMP9 in tumor
Non-dual high expression of MIF and MMP9 in Grade I tumor
Dual high expression of MIF and MMP9 in Grade I tumor
Log Rank test P=0.001
Log Rank test P=0.005
0
20
40
60
80
100
120



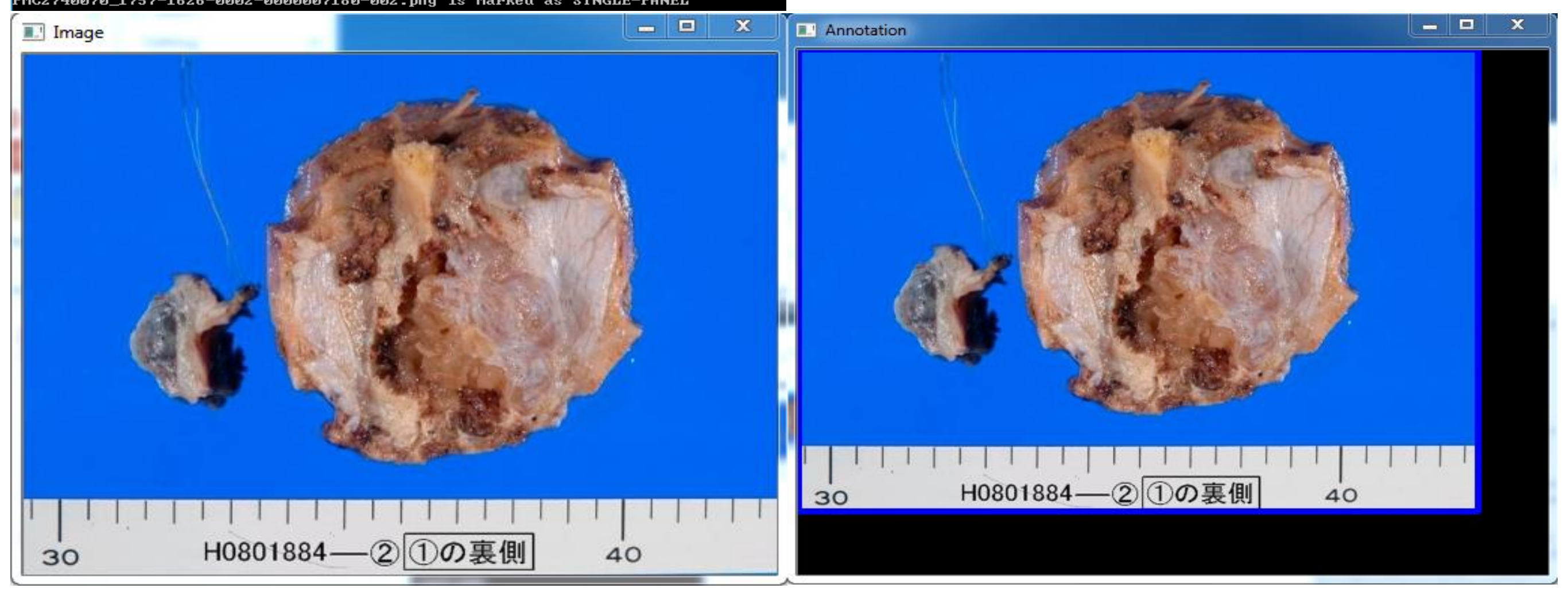
```
7: Visualize Annotation for 11-completed\PMC3558716_ijmsv10p0276g02.png
The style is STITCHED
Press n, go to the next figure
Press p, go to the previous figure
Press 1, mark it as SINGLE-PANEL
Press 2, mark it as MULTI-PANEL
Press 3, mark it as STITCHED-MULTI-PANEL
PMC3558716_ijmsv10p0276g02.png is Marked as STITCHED-MULTI-PANEL
```

(D) The left image is the original image. The right image is the style annotated image. This is an example of a stitched-multi-panel figure, because there is no distinct boundary for each of the individual panels. This example also shows how the annotations look with more panels, and how the colors identifying the panels and labels stay different.

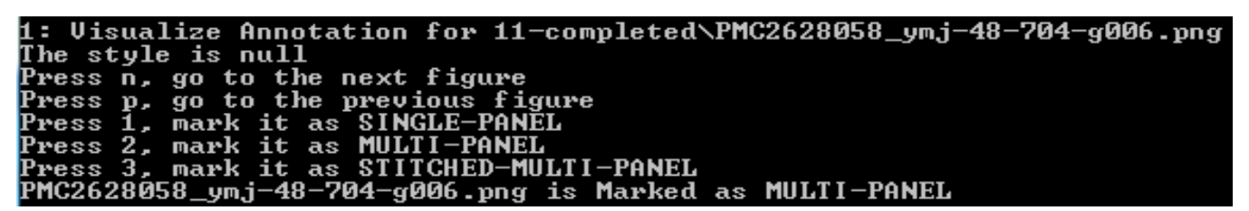


```
4: Visualize Annotation for 11-completed\PMC2740070_1757-1626-0002-0000007180-00
2.png
The style is null
Press n, go to the next figure
Press p, go to the previous figure
Press 1, mark it as SINGLE-PANEL
Press 2, mark it as MULTI-PANEL
Press 3, mark it as STITCHED-MULTI-PANEL
PMC2740070_1757-1626-0002-0000007180-002.png is Marked as SINGLE-PANEL
```

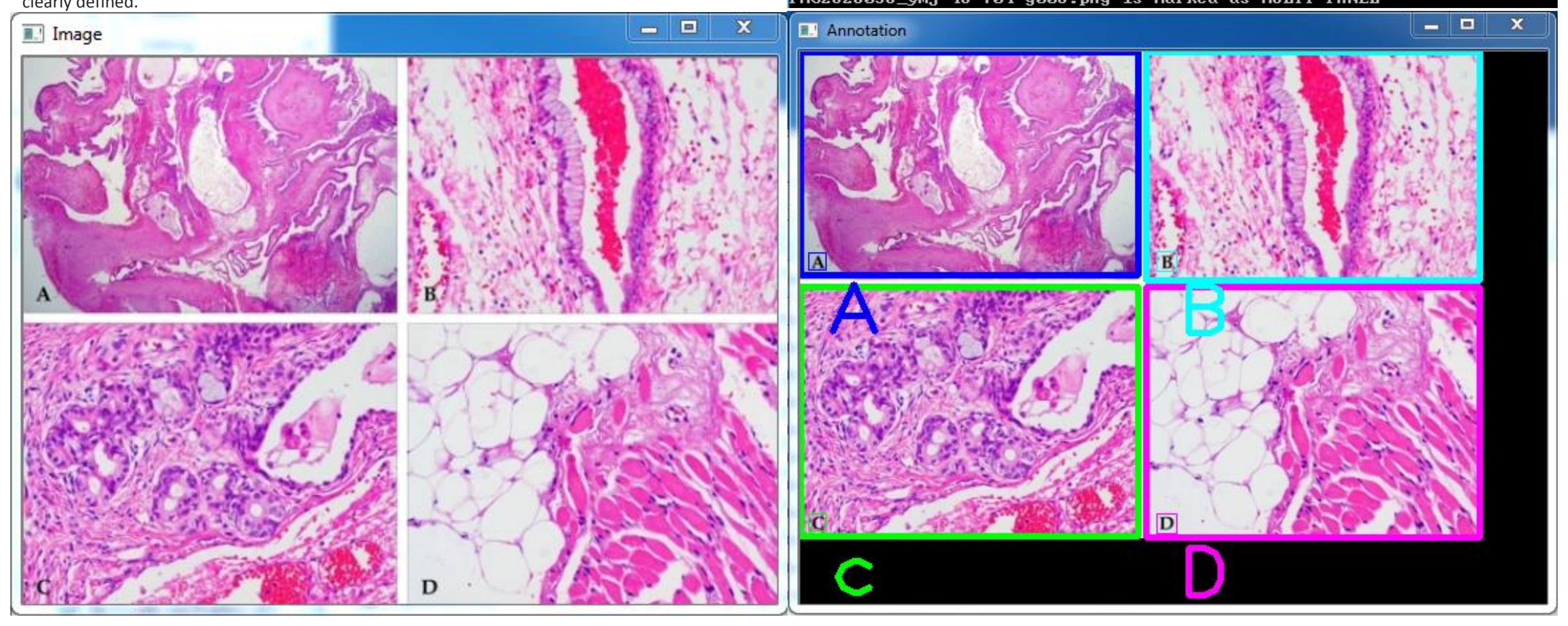
(E) This is the same example as the single-panel figure before, but this shows what the style annotation looks like for a single-panel figure. The bottom left image is the original figure before paneling, and the bottom right image is after being style annotated. As a result of it being a single-panel figure there are no panel labels.



(F) The left image is the original image after being downloaded and before being paneled from part one. The right image is after the annotation program is run. It matches the panels and labels in order to make sure there are not still errors. After checking to make sure everything is the same, a proper label is given to the figure. This figure is multi-panel because the individual panels are clearly defined.



```
1: Visualize Annotation for 11-completed\PMC2628058_ymj-48-704-g006.png
The style is null
Press n, go to the next figure
Press p, go to the previous figure
Press 1, mark it as SINGLE-PANEL
Press 2, mark it as MULTI-PANEL
Press 3, mark it as STITCHED-MULTI-PANEL
PMC2628058_ymj-48-704-g006.png is Marked as MULTI-PANEL
```



PROBLEMS

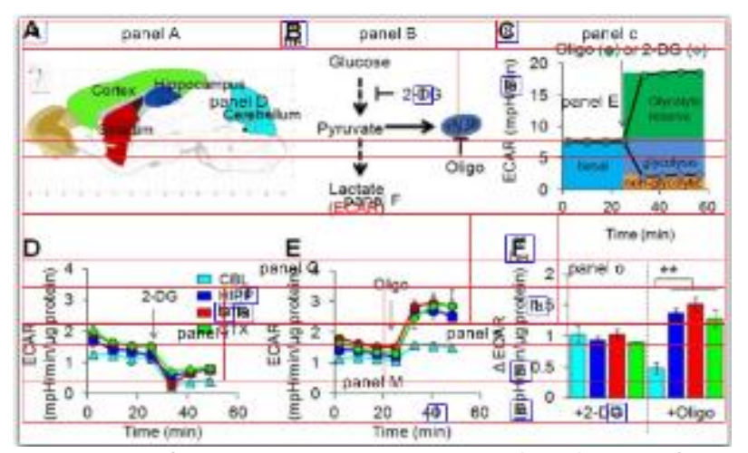
In Panel Collection and Annotation, the main problem occurred with the PanelSegEval.jar algorithm because it could not handle certain figure types. Some of the problems are in the following types of images:

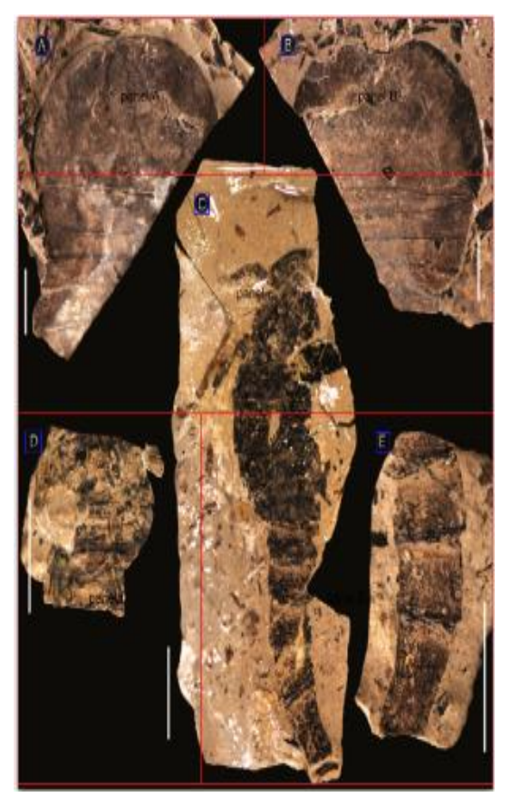
- Graphs
- Any with data charts
- Images with extra words or random letters
- Multi-panel figures with overlapping parts

As a result, the figures would have some of the following problems:

- Extra panel boxes
- Extra label boxes
- Wrong labeling on letters
- No labels when they were not needed
- Labels when none were needed

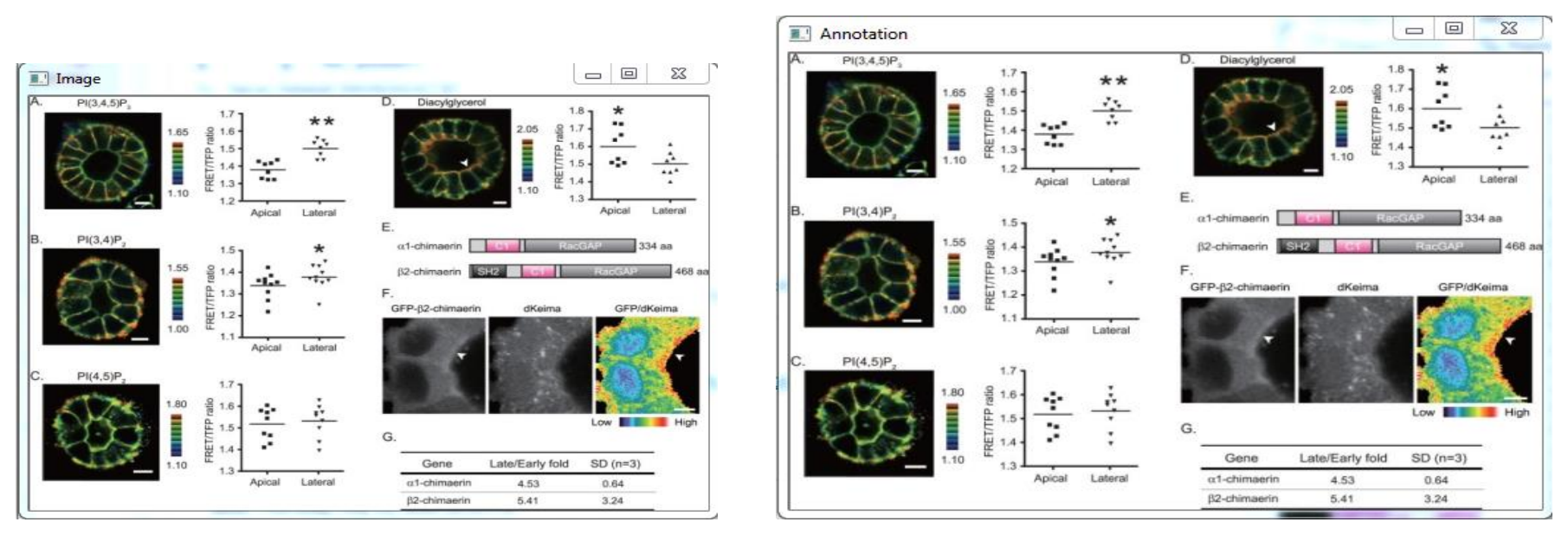
See images A, B, and C in Part One for examples, as well as to the right.





In Verification, the main problem was the amount of errors that were recorded in the txt file, because most of annotation issues were not found when the figures were rechecked.

In Style Assignment, the main issue was not having the style annotation boxes and labels appear on the annotated image.



The far left image is the figure after downloading, and the middle image is the figure after annotating. This one is identified as a Stitched-Multi-Panel, because there are multiple parts even though they are not in bounding boxes.

SUMMARY

Following this workflow pipeline, we are able to collect and annotate figures efficiently. Over a period of 7 weeks, 10,262 figures are collected and ground-truth annotated by one person. With this larger dataset, a lot more rigorous evaluation can be conducted and algorithms relying more on machine learning instead of hand-crafted rules can be researched. We believe this dataset is valuable to the future R&D of figure processing.

REFERENCES

1. D. You, S. Antani, D. Demner-Fushman, V. Govindaraju, G.R. Thoma, Detecting Figure-Panel Labels in Medical Journal Articles Using MRF, 967-971, Int'l Conf. on Document Analysis and Recognition, 2011
2. E. Apostolova, D. You, Z. Xue, S. Antani, D. Demner-Fushman and G.R. Thoma, Image retrieval from scientific publications: Text and image content processing to separate multipanel figures, Journal of the American Society for Information Science and Technology, 64:5, 893–908, 2013
3. K.C. Santosh, S. Antani, G.R. Thoma, Stitched Multipanel Biomedical Figure Separation, 54-59, IEEE 28th Int'l Symposium on Computer-Based Medical Systems, 2015