**PANCREAS VASCULOPATHY DATA PROCESSING January 2021**

The aim of this work is to calculate the degree of arterial occlusion in a series of 224 pancreas tissue slides and to derive some statistics from these measurements.

The initial annotation to generate this information is carried out using Leica Aperio ImageScope software. The output from this annotation stage is then processed by text editing in Emacs, re-formatted using bespoke Perl scripts and the final statistical analysis carried out in Excel.

There are three stages to this work as follows:

ANNOTATION: generating the raw data

ANALYSIS\_STAGE\_1: calculates the percentage occlusion of four blood vessels in each of the 224 tissue slices

ANALYSIS\_STAGE\_2: calculates the statistics for the four blood vessels in each tissue slice

**ANNOTATION**

On each tissue slide, four blood vessels are annotated. For each blood vessel, three lines are drawn in order to define (a) the outer blood vessel wall, (b) the inner vessel wall and (c) the boundary of any occluding material in the lumen of the blood vessel. The software calculates the areas enclosed within each of these three lines and puts the data into an Excel file. From these area figures, the percentage occlusion of the blood vessel can be determined.

Examples of an annotated blood vessel can be seen in file ‘Annotation\_Example.jpg’ and the figures determined from this can be seen in the file ‘Annotation\_Example\_Output.xlsx’.

The format of the ImageScope Excel files is not ideal for calculating the % occlusion and the subsequent statistics. The subsequent steps aim to convert this output into a format more amenable to such analysis. This is done in two stages.

**ANALYSIS\_STAGE\_1**

The overall aim in STAGE\_1 of the analysis is to determine the percentage occlusion for four blood vessels on each slide.

To do this, the Excel files are converted to \*.txt files and combined into a single text file. A Perl script then takes this file and combines the three lines of data for each blood vessel onto a single line. The output file is opened in Excel, edited to remove unwanted columns and the % occlusion for each blood vessel is calculated. The final Excel file for this stage (Vasculopathy\_All\_Stage\_1\_v3.xlsx) is shown and each line shows the % occlusion for a single blood vessel.

The steps in Stage 1 analysis are as follows:

1. Rename \*.xls.xlsx files to \*.xlsx

2. Convert the \*.xlsx files to \*.csv files.

3. Concatenate the \*.csv files

4. Edit the concatenated file

5. Run the 'vasculopathy.pl' script

6. Open output file in Excel and calculate vasculopathy scores

**# STEPS 1-5 ARE DONE IN UNIX/LINUX ENVIRONMENT ON CSD3 OR EQUIVALENT**

**1. RENAME \*.xls.xlsx files to \*.xlsx**

Open Thunar file manager in CSD3 to launch bulk rename utility:

thunar -B

Add files

Select -> Open

In boxes:

Search & Replace

Name and Suffix

Search for: .xls.xlsx

Replace with: .xlsx

Rename files

# This step shouldn't have been necessary - the \*.xls.xlsx nomenclature is an annotation error

**2. CONVERT \*.XLSX FILES TO \*.CSV FILES**

Install 'csvkit' software on Linux/Unix environment (see notes)

**a). USING EMACS**

In Linux:

ls > newfile.txt

Open 'newfile.txt' in Emacs and edit each line to:

'~/.local/bin/in2csv filename.xlsx > filename.csv'

Save as 'newfile\_edited.txt' and then 'source' the file:

source newfile\_edited.txt

-> Converts all \*.xlsx files to \*.csv files

**b). ALTERNATE METHOD:**

Shell script in Linux (Bash shell):

for i in \*.xlsx; do libreoffice --headless --convert-to csv "$i"; done

# This works on locally-installed Linux that has LibreOffice installed

- May not work on CSD3

**3. CONCATENATE THE \*.CSV FILES**

In Linux:

ls > \*.csv file\_list.txt

# Puts all \*.csv filenames into single file (file\_list.txt)

Open 'file\_list.txt' in Emacs and edit each line to:

cat file\_1.csv > combined\_file.txt

cat file\_2.csv >> combined\_file.txt

cat file\_3.csv >> combined\_file.txt

etc...

# Note use of '>' for first line, thereafter '>>'

# '>' creates the file, '>>' appends to the end of the file you've just created

Save as 'file\_list\_edited.txt' and then 'source' the file:

source file\_list\_edited.txt

-> All \*.csv files are concatenated into 'combined\_file.txt'

**4. EDIT CONCATENATED FILE**

a. Open 'combined\_file.txt' in Emacs and removed header lines (begin with 'Sample...')

Save as 'combined\_file\_edited.txt'

b. Move 'combined\_file\_edited.txt' to same directory as Perl script (vasculopathy\_stage\_1.pl)

Rename to 'input\_file\_stage\_1.txt'.

**5. RUN PERL SCRIPT (vasculopathy\_stage\_1.pl)**

Takes concatenated \*.txt file as input. Puts the three lines of data for each blood vessel onto a single line. This makes it easier to edit in Excel in order to calculate the % occlusion figures.

Run script 'vasculopathy\_stage\_1.pl'.

Output file is 'outfile\_stage\_1.txt'.

Rename 'outfile\_stage\_1.txt' to 'outfile\_stage\_1.csv'.

Transfer 'outfile\_stage\_1.csv' back to Windows laptop

**6. CALCULATE VASCULOPATHY SCORES**

Calculate % occlusion for all blood vessels in each sample

Open 'outfile\_stage\_1.csv' in Excel

- Delete unneeded columns # Just left in as a visual check that all is as it should be

- Calculate 'Occluded\_Area' (= total lumen area - free lumen area)

- Calculate '%\_Occlusion' (= (occluded lumen area / total lumen area) x 100%)

- Save as 'Vasculopathy\_All\_Stage\_1.xlsx'

**- Final file, after edits, is 'Vasculopathy\_All\_Stage\_1\_v3.xlsx'**

#This shows the vasculopathy scores for each blood vessel in a given tissue slice.

#There is a separate line for each blood vessel.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**ANALYSIS\_STAGE\_2**

This stage combines the occlusion data from each of the four blood vessels in a given tissue slice into a single line, in order to calculate the aggregate vasculopathy scores for each tissue section overall. The process is broadly the same as what was done in the Stage\_1, but with some differences. The steps in Stage\_2 analysis are:

1. Edit the output Excel file from Stage 1 ('Vasculopathy\_All\_Stage\_1\_v3.xlsx')

2. Move file to Unix/Linux environment

3. Rename input file

4. Run Perl script

5. Open output file in Excel and calculate vasculopathy scores

**1. EDIT EXCEL FILE FROM STAGE 1**

Starting file: Vasculopathy\_All\_Stage\_1\_v3.xlsx

Make a copy and move it to the Analysis\_Stage\_2 directory

Rename as 'stage\_2\_input.xlsx'

Edit to remove non-required data

Keep only first and last columns

Copy last column & paste as 'values'

Delete header line

Save as 'stage\_2\_input\_edited.csv'

-> this is input for script 'vasculopathy\_stage\_2.pl'

**2. MOVE THESE FILES TO UNIX/LINUX ENVIRONMENT**

stage\_2\_input\_edited.csv

vasculopathy\_stage\_2.pl

**3. RENAME INPUT FILE FOR PERL SCRIPT**

Rename 'stage\_2\_input\_edited.csv' to 'infile\_stage\_2.txt'

mv stage\_2\_input\_edited.csv infile\_stage\_2.txt

# This is the filename the script looks for

# Might work with \*.csv files - haven't tested

**4. RUN PERL SCRIPT ('vasculopathy\_stage\_2.pl)**

Each tissue section has four annotated blood vessels. Each one of these is on a separate line, so have to work with groups of four lines per tissue slice. For each group of four lines, the script takes the first and last fields of the first line (sample ID plus first % Occlusion figure) and concatenates them with last field (% Occlusion) from remaining three lines in the group.

Script output is 'outfile\_stage\_2.txt'

Move back to laptop and rename as 'stage\_2\_output.csv'

**5. CALCULATE VASCULOPATHY SCORES PER TISSUE SECTION**

Calculate overall % occlusion for each tissue section

Open 'stage\_2\_output.csv' in Excel

Edit as required

Do stats analysis

**Save as 'Vasculopathy\_All\_Final.xlsx'**

**NOTES**

i). Convert \*.xlsx files to \*.csv files

See this web page:

https://stackoverflow.com/questions/10557360/convert-xlsx-to-csv-in-linux-with-command-line

Go from entries on there, to:

https://csvkit.readthedocs.io/en/1.0.5/tutorial/1\_getting\_started.html

Install 'csvkit' software

ALTERNATELY: install 'xlsx2csv' software

Other methods discussed in above Stack Overflow web page.