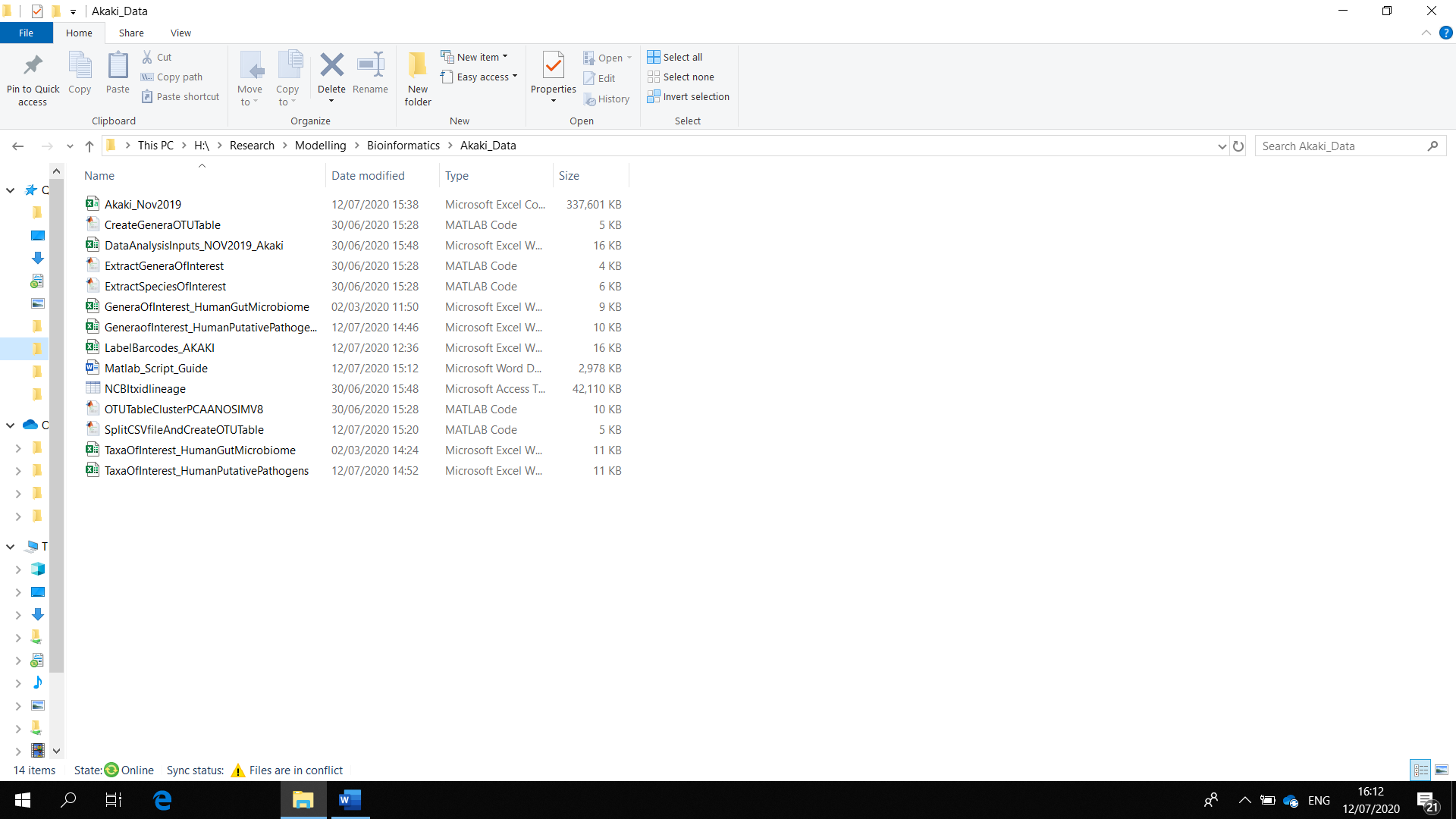
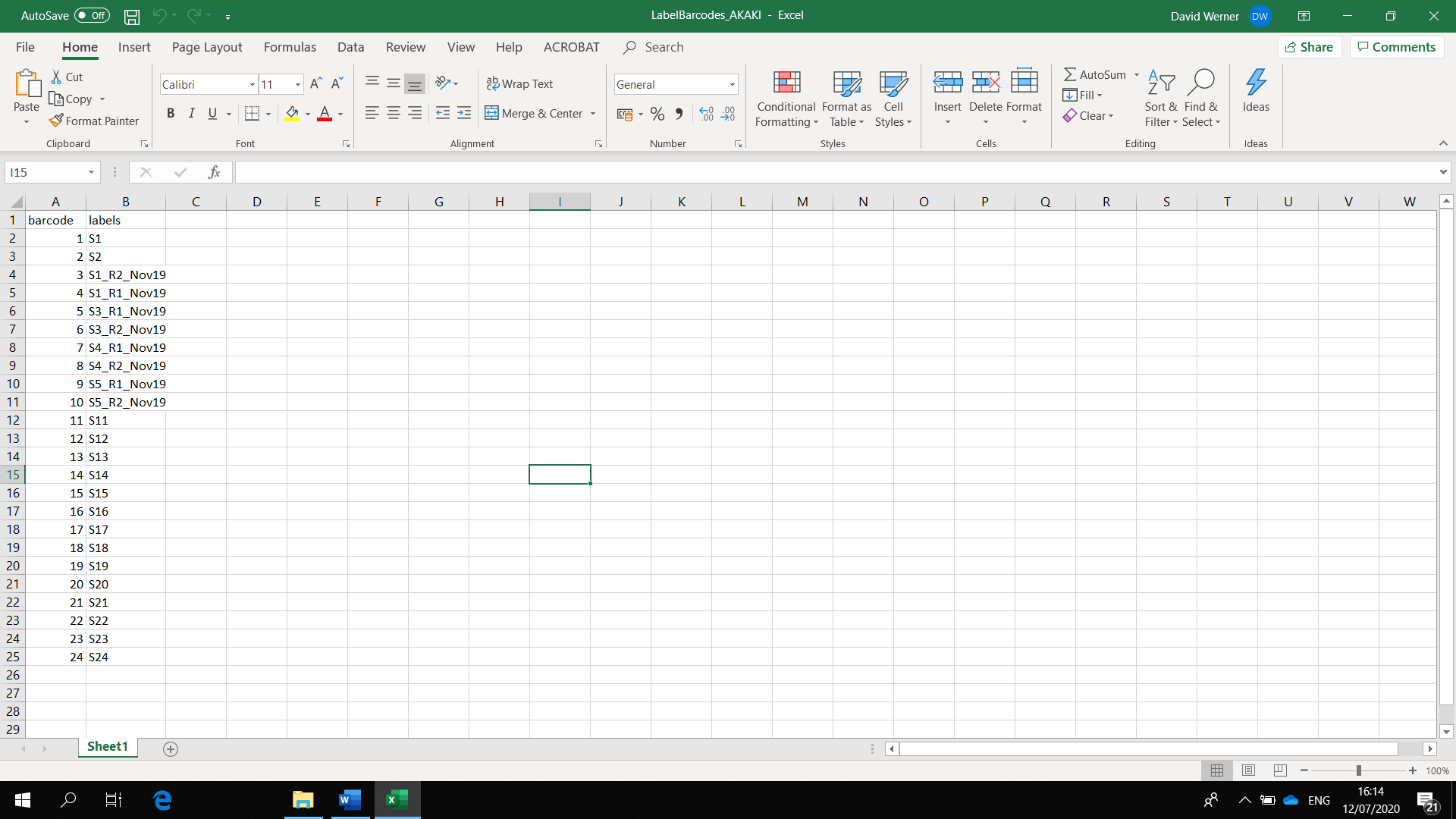
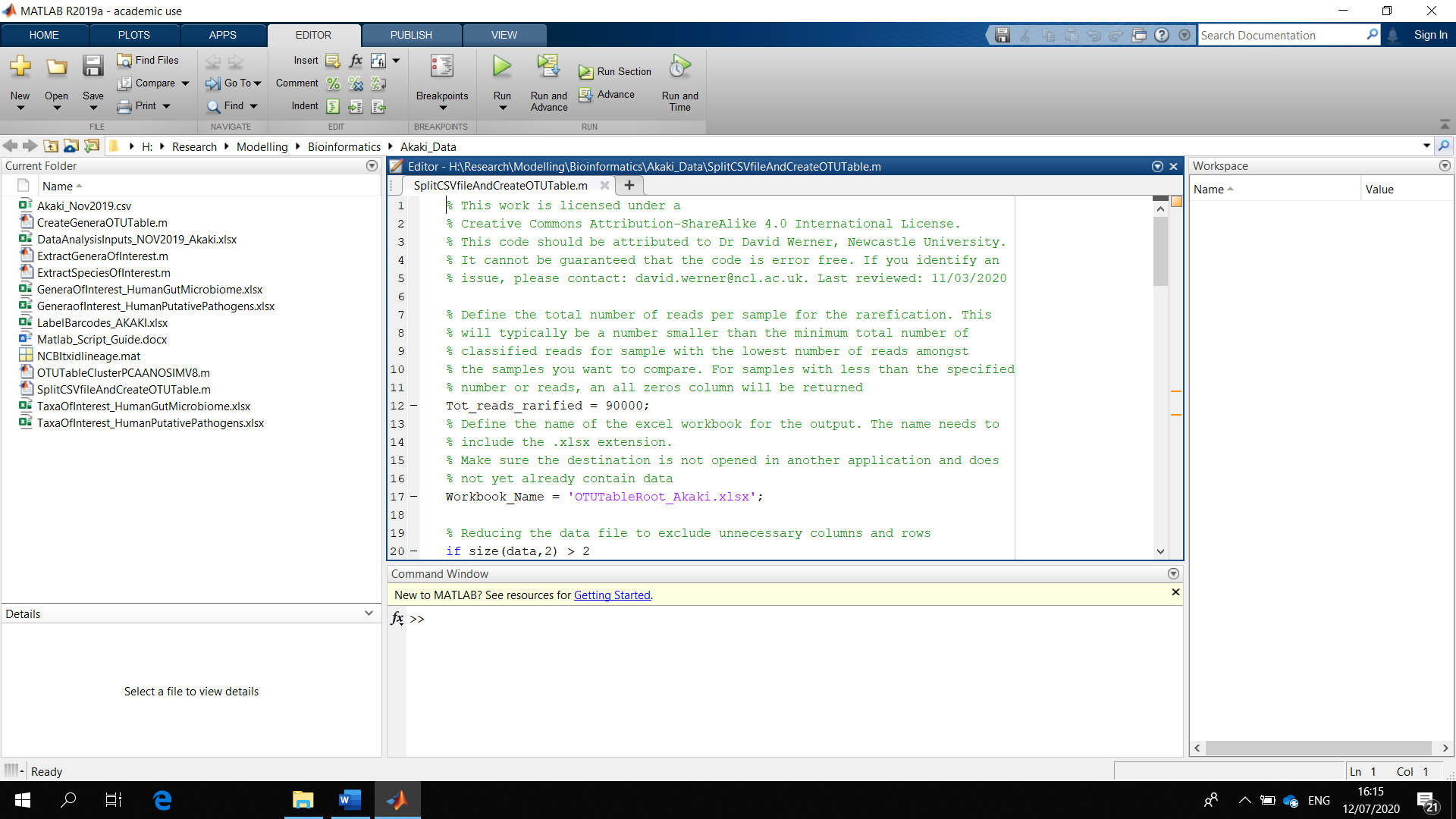
It is recommended to close all other applications and reboot your computer before processing the data to free up as much memory as possible. Create a project folder which contains the relevant files: The Matlab codes, the NCBI library, your excel files with the barcode info, data analysis inputs, taxa and genera of interest, etc (prepared following the format of the templates provided), and all the CSV files.



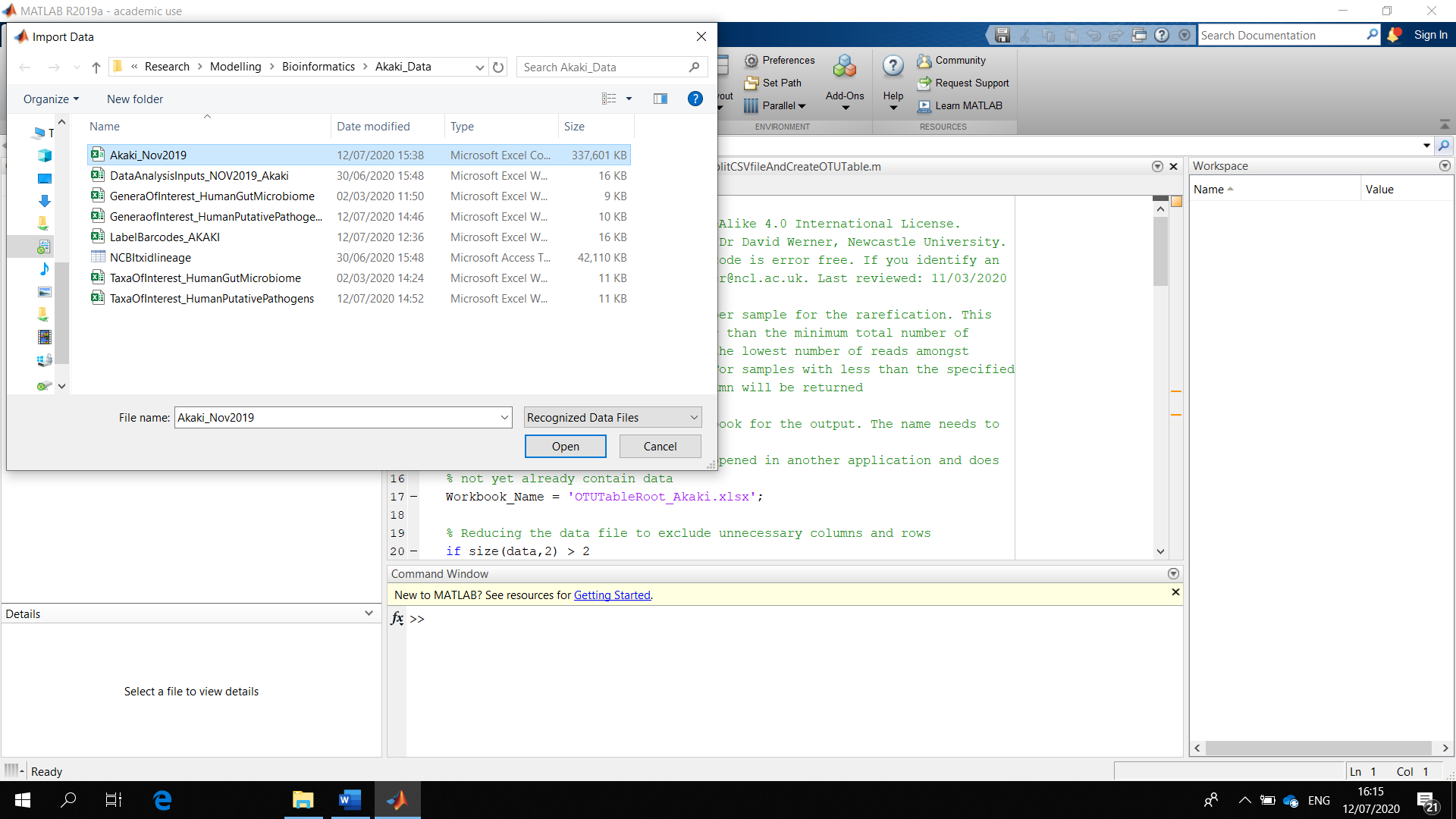
Edit the LabelsForBarcodes file to provide a unique sample ID for each sample. If you have less than 24 barcodes, you still need to provide labels for the other barcodes in case the data processing matched some of these barcodes by mistake. Close the excel file when you are done.



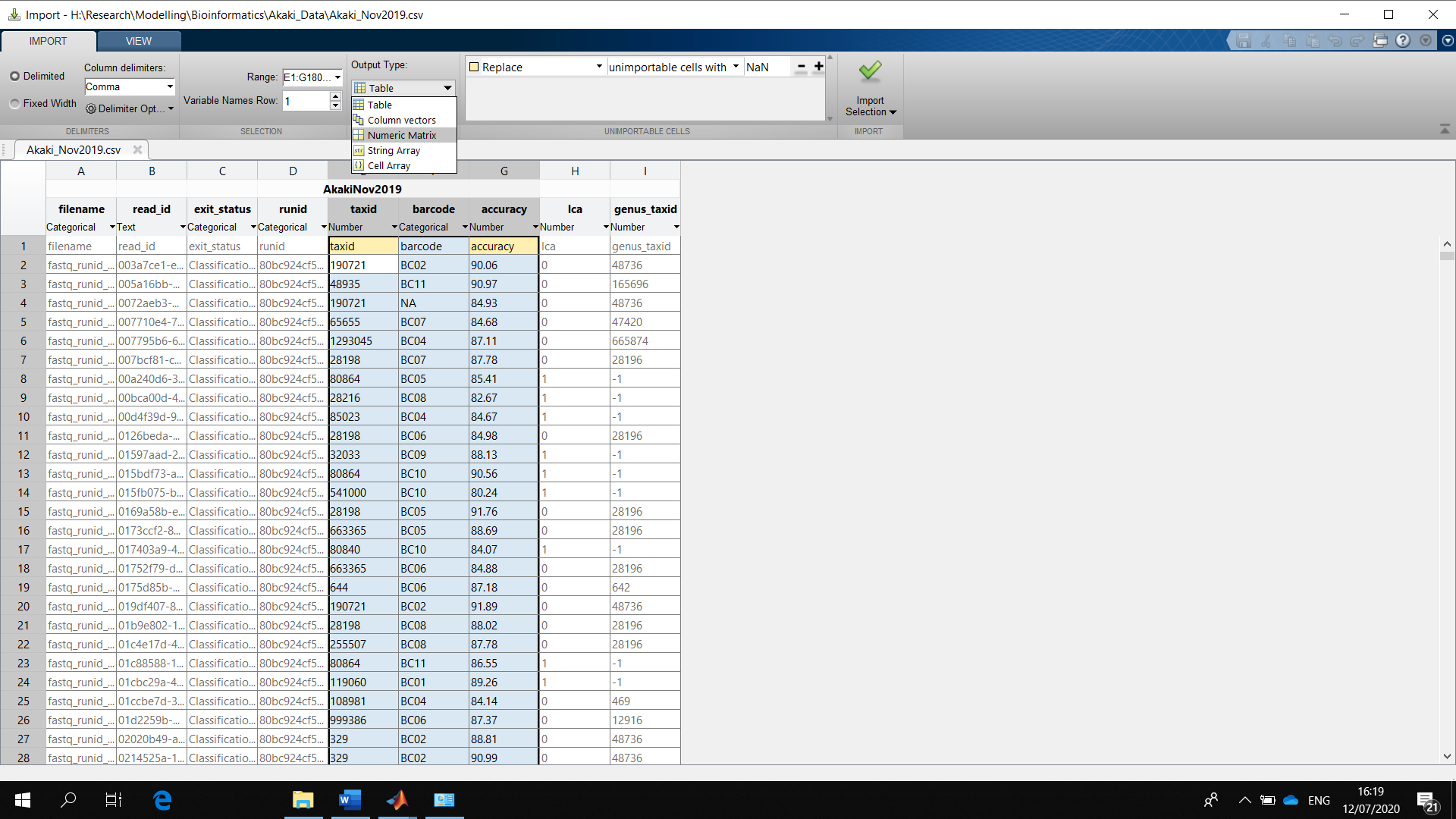
Set the Matlab path to the relevant folder and open the file “SplitCSVfileAndCreateOTUTable” by double clicking on it:



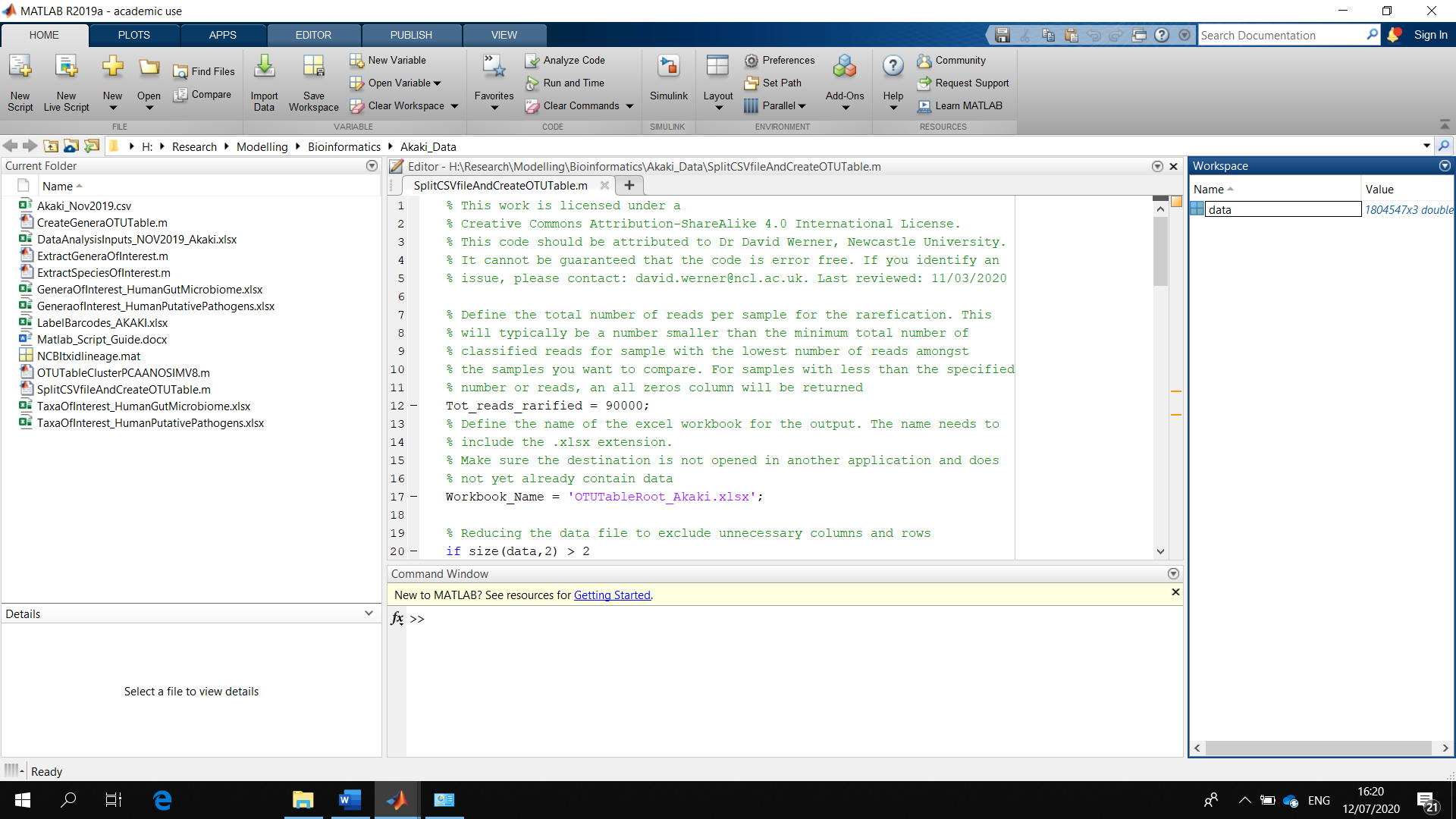
In the Home tab, use the “Import data” function in the Home tab to open the CSV file you want to process. A warning will show that it is a large file. Depending on your computer it may take a while to open.



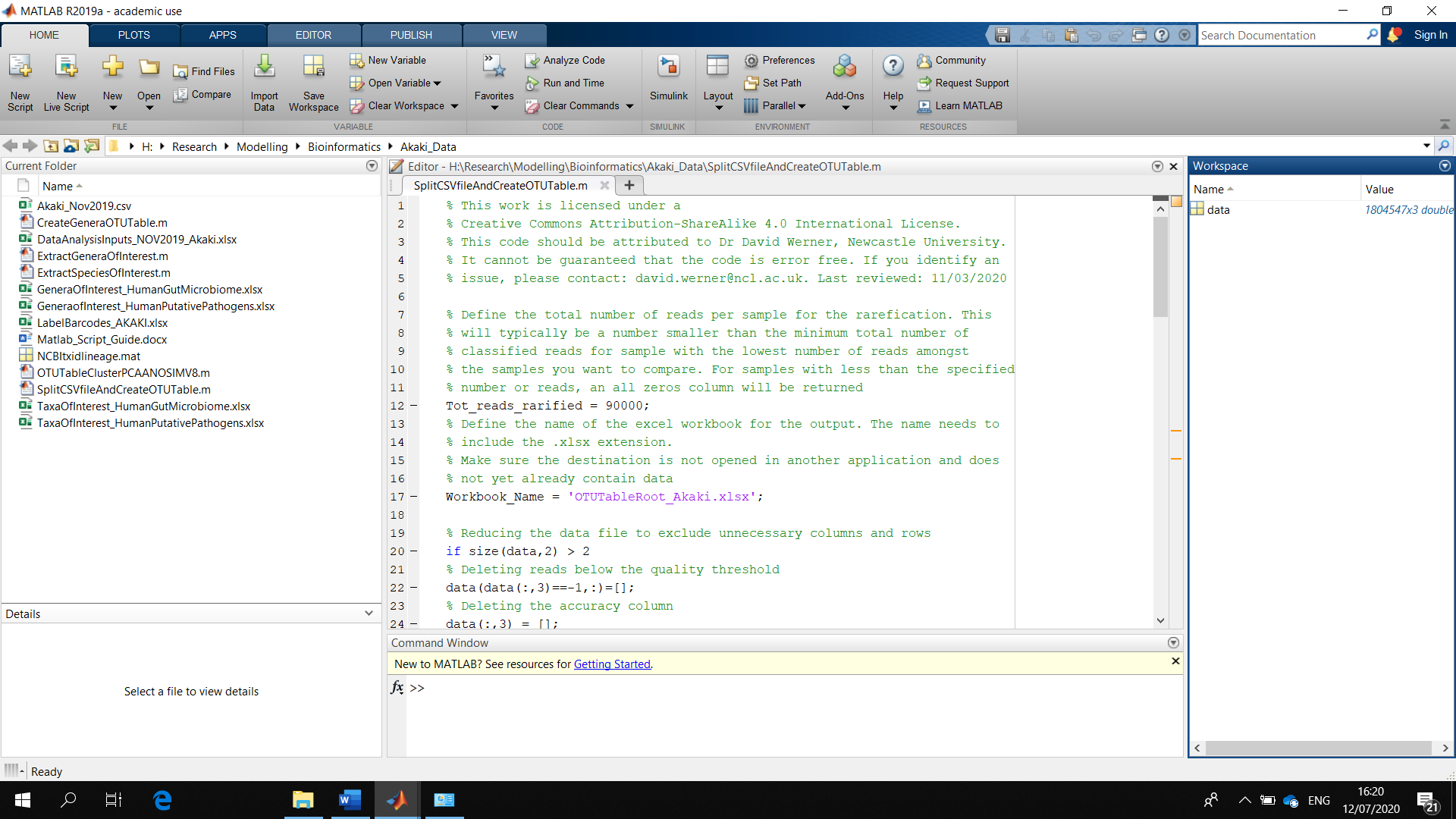
Highlight the columns taxid, barcode, and accuracy (by swiping over the top row, E,F,G, in the example below) to select them for import. Import the file as “Numeric Matrix” by choosing that option from the drop-down menu. Then click “Import Selection”. Another warning will show about the size of the file. Click yes. Depending on your computer it may take a while to import. Once the warning is gone, the file has been imported and you can close the import window.

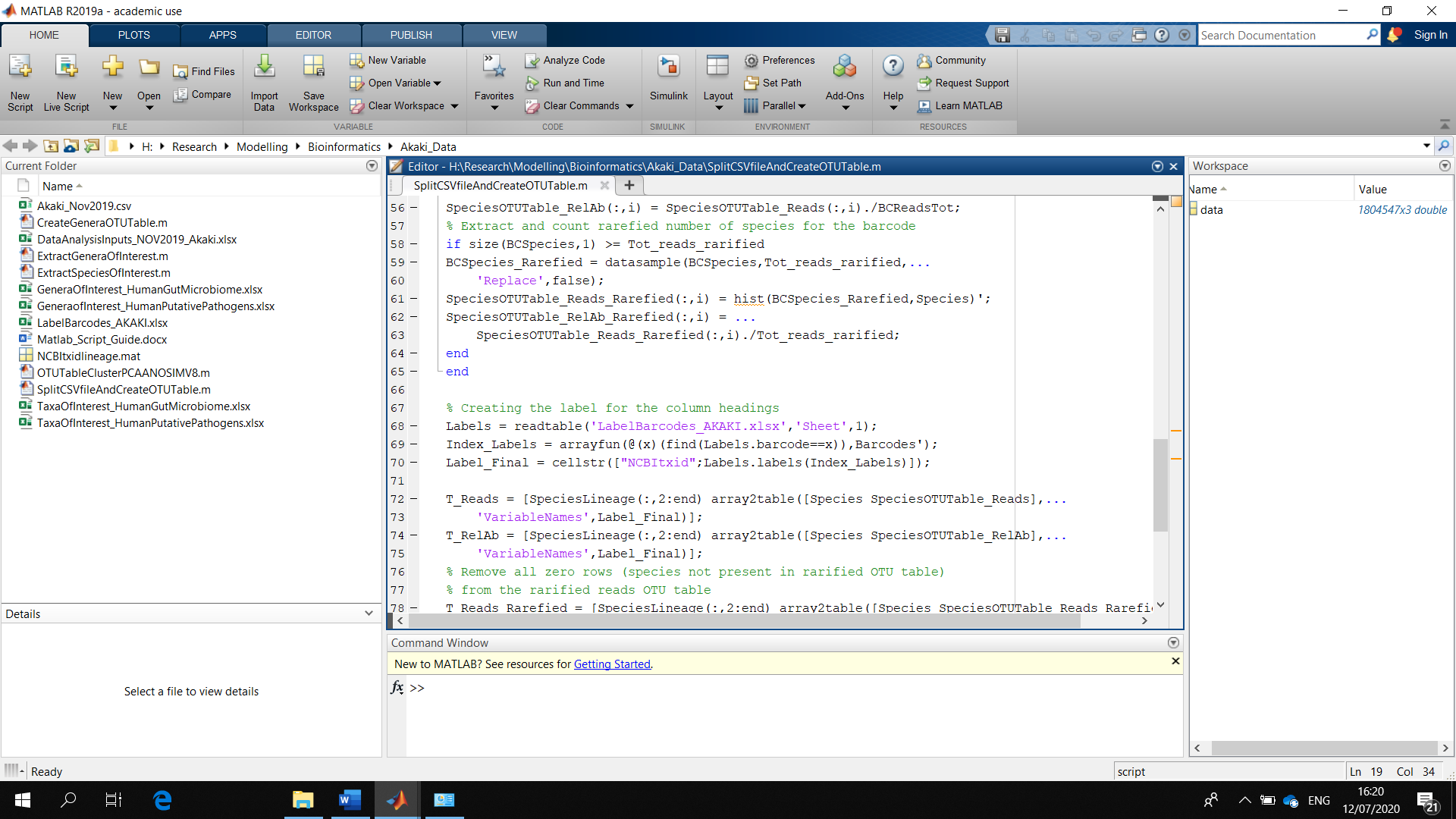


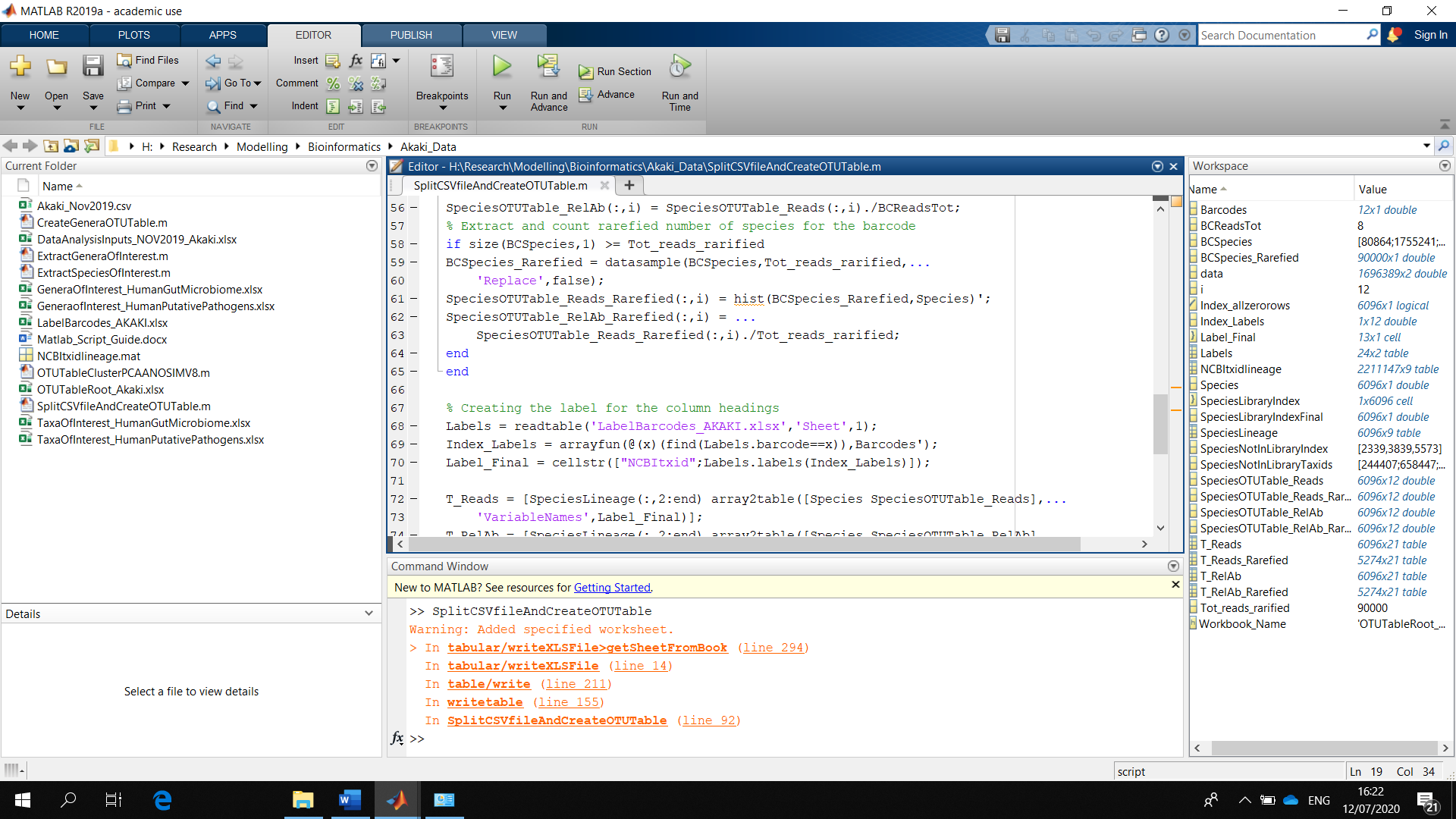
Rename the imported file “data” by clicking on the file name in the workspace



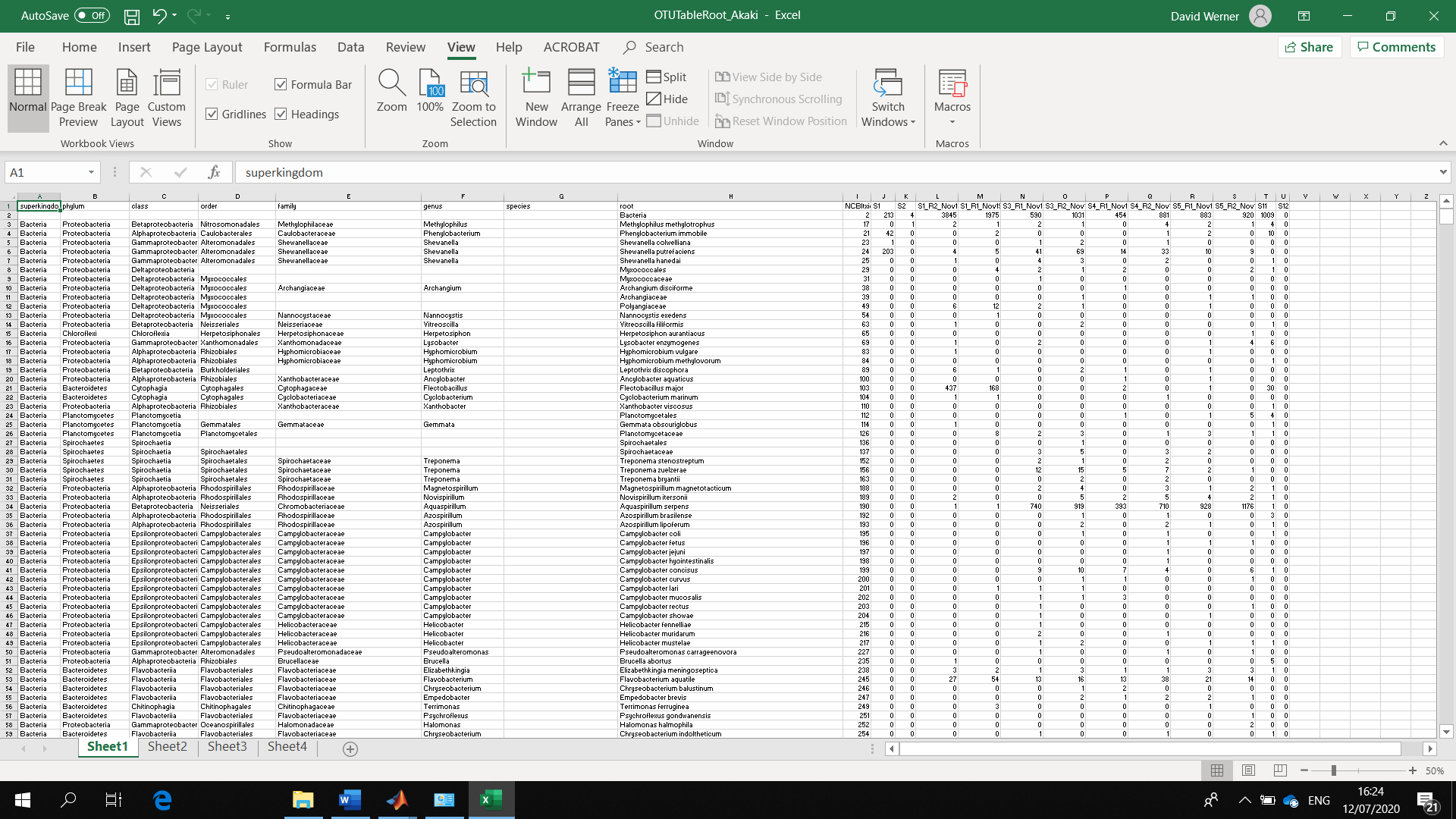
In the Editor tab, edit the file to define the total number of reads for the rarefied output, and the name of the output file (in the example line 17 OTUTableRoot\_Akaki\_Nov2019.xlsx, the “.xlsx” at the end is important). For the rarefaction (line 12), you will normally choose a number of reads less than the minimal number of reads above the quality threshold across all the samples you are analyzing in your project. However, to include rare species, you want to have at least 10,000 reads, better more. If you have a sample (barcode) with very low number of reads, you may chose a higher number for rarefaction, and an all zeros column will be returned for the samples with less reads in total. You also need to provide the labels for the barcodes (in the example line 68 LabelBarcodes\_AKAKI.xlsx) In the Editor tab, click on the green “run” arrow to execute the commands. You may have to click on the big window in the middle to see the editor tab. While the code is running, it will say “Busy” in the bottom left corner of the window. This may take a while.



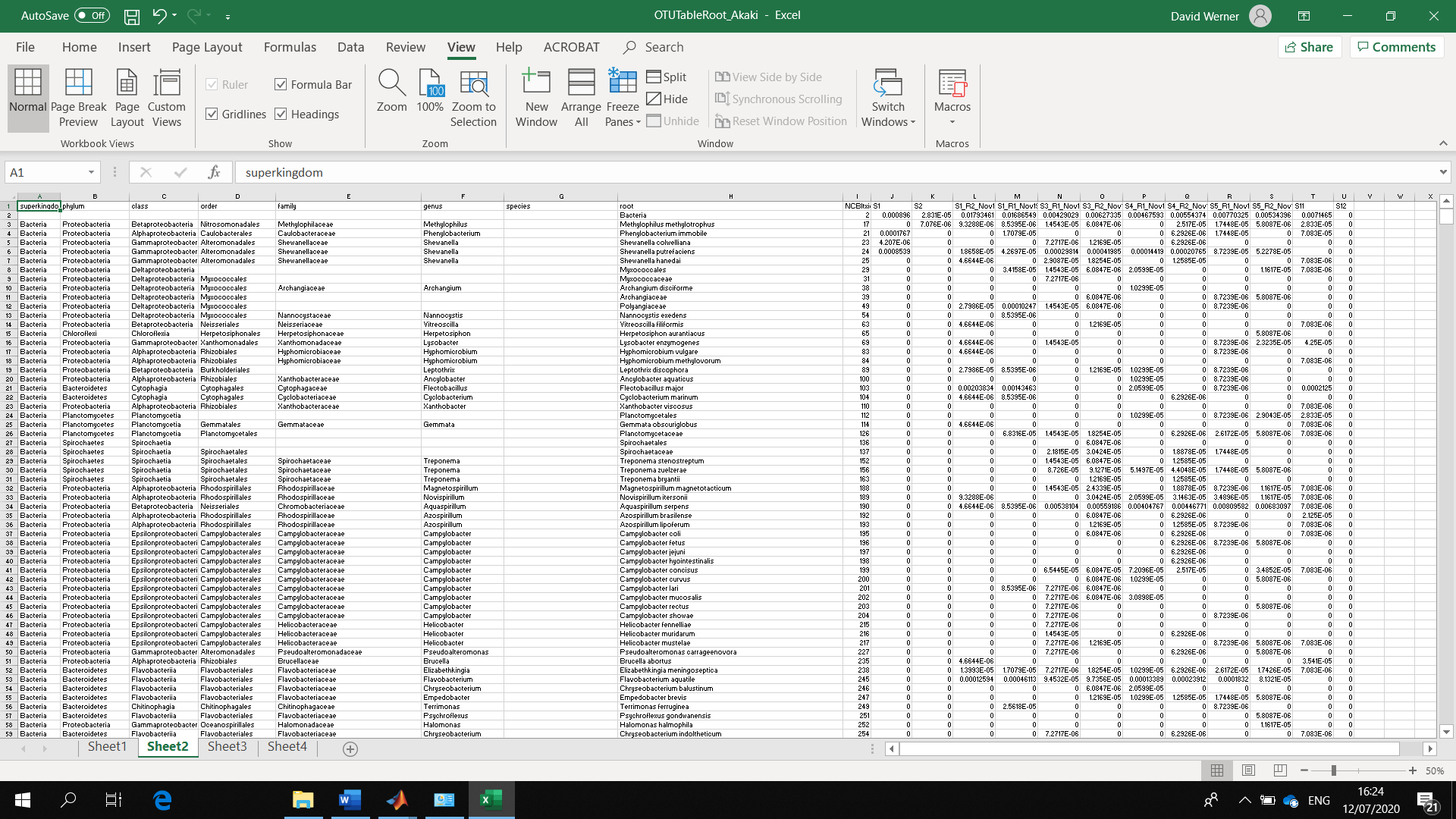


The output files will be created. The main one in this example is “OTUTableRoot\_S3.xlsx” 

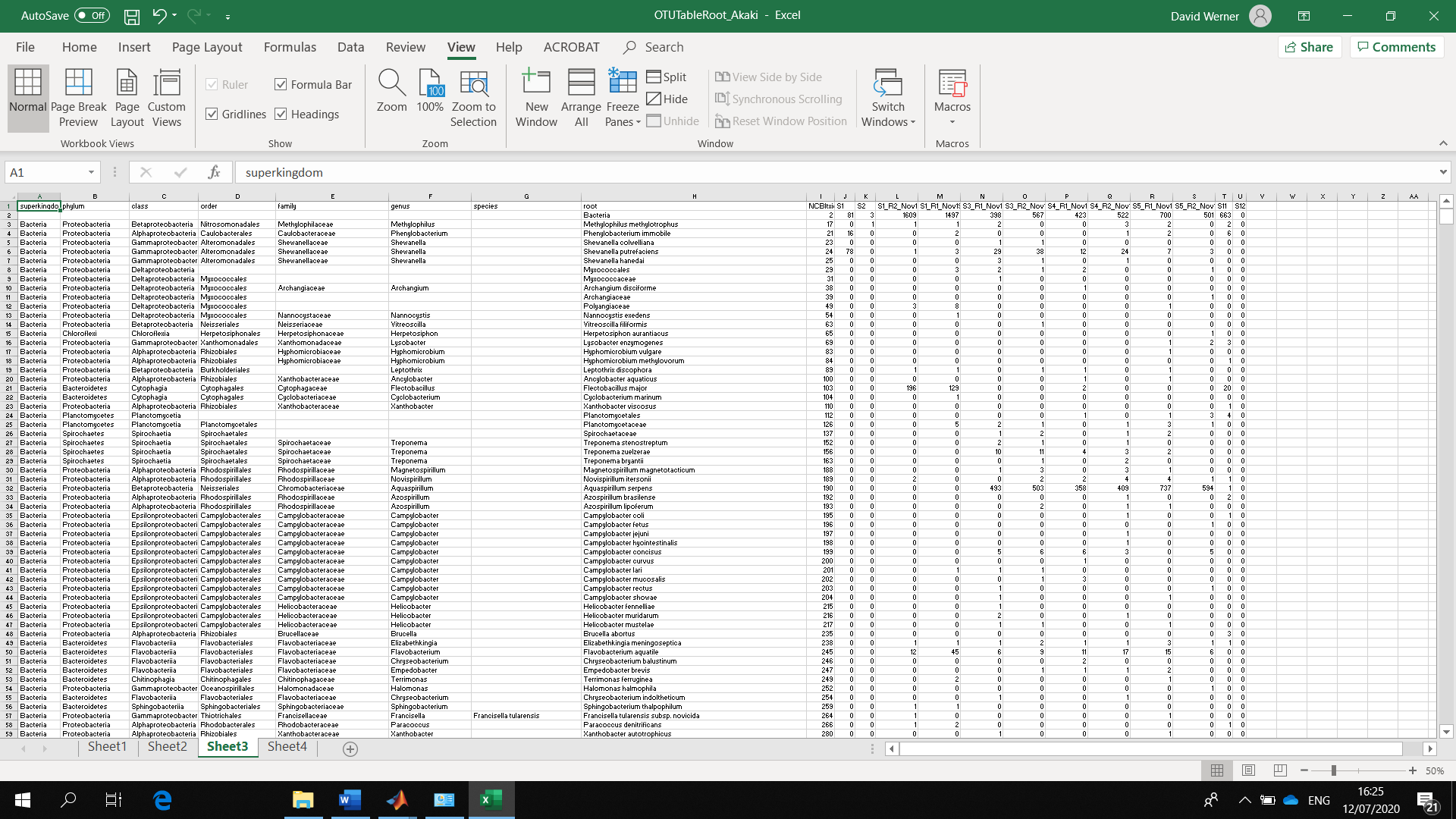
You can open the excel files from your folder. The first spreadsheet contains the complete OTU table with the lineage and the numbers of reads. Because of the library set-up, the species information is often only found in the “root” column.



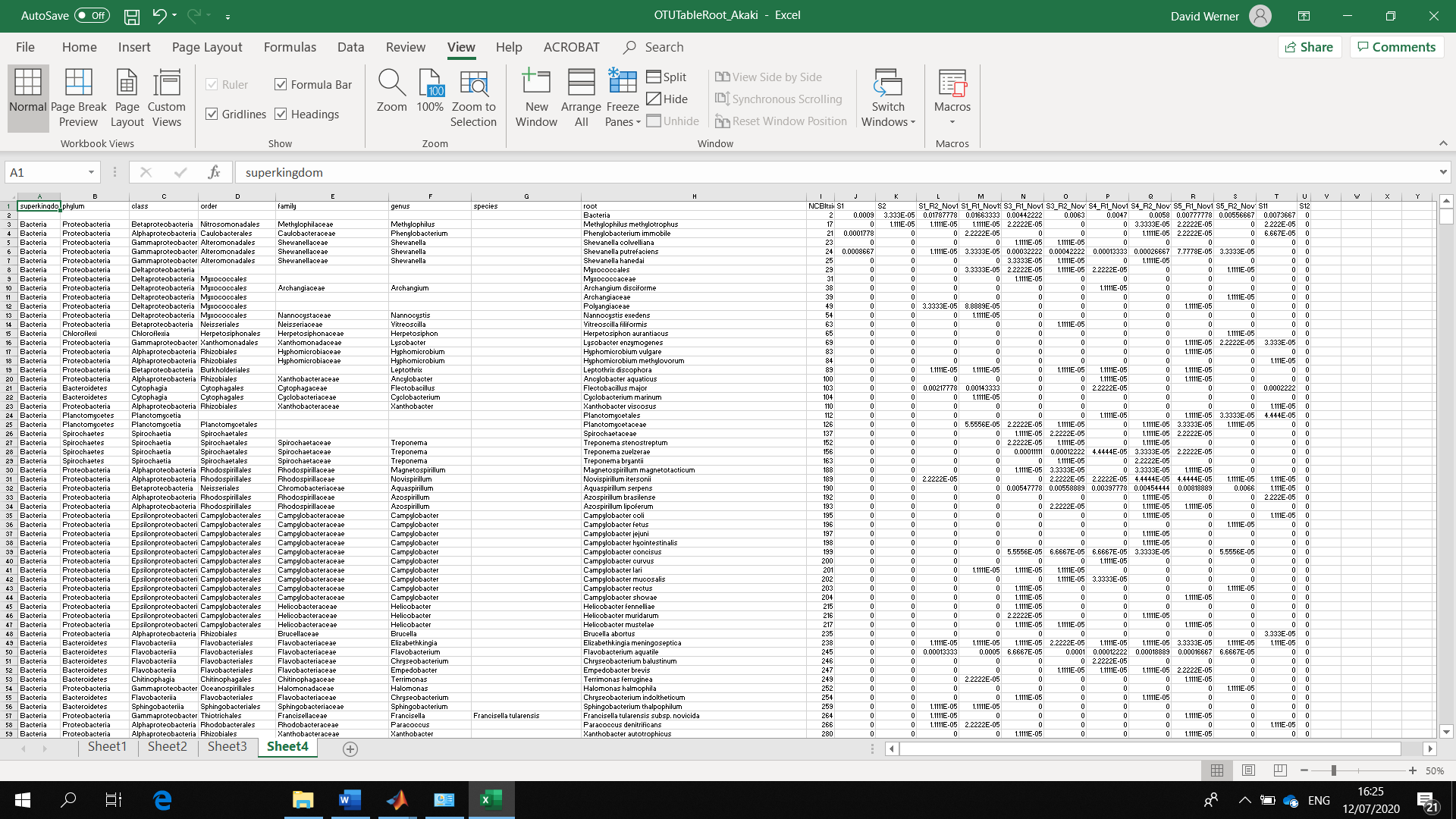
The second spreadsheet contains the relative abundance of reads (number of reads for each OTU divided by the total number of reads in the sample.



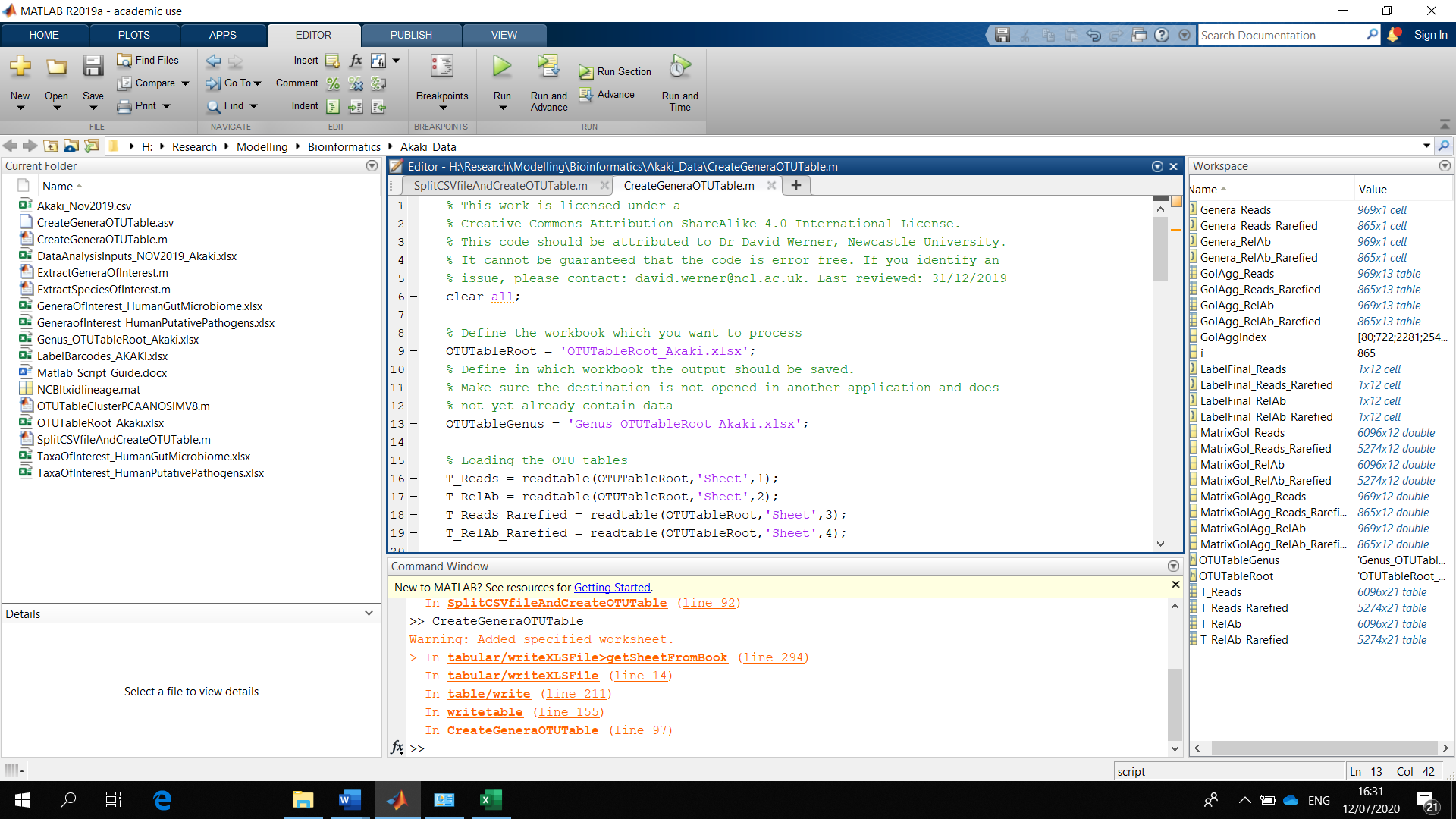
The third spreadsheet contains the complete OTU table with the lineage and the numbers of reads for the rarefied samples. In this spreadsheet the total number or reads for each sample should be the same, making them more comparable.



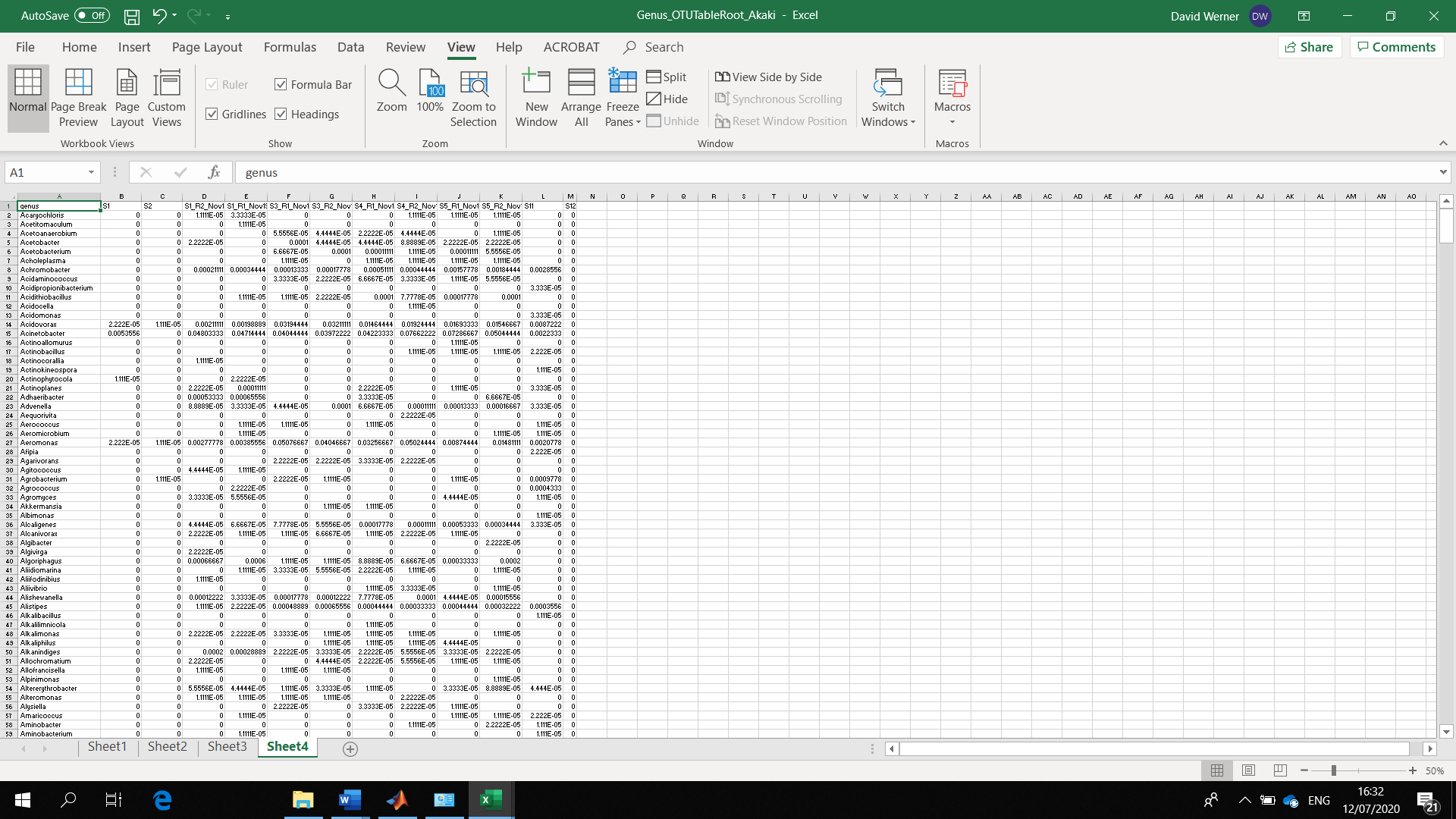
The fourth spreadsheet contains the complete OTU table with the lineage and the relative abundance of reads for the rarefied samples. In this spreadsheet the sum for each sample/column should equal 1. The rarefied relative abundance table is the data which is typically used in analysis of the similarity between samples (cluster analysis, PCA, ANOSIM, etc).



You can generate an OTU table grouping reads at genus level for reads which have been classified to at least this level. This is recommended, as MinION read errors make species level data unreliable, although it may be interesting for exploration. Open the CreateGeneraOTUTable file. Specify which OTU table with root level data you want to process (line 9), and the name of the genus level output file (line 13). Click run. This may take a while.



The most interesting data is in sheet 4, which is again the rarefied relative abundance data.



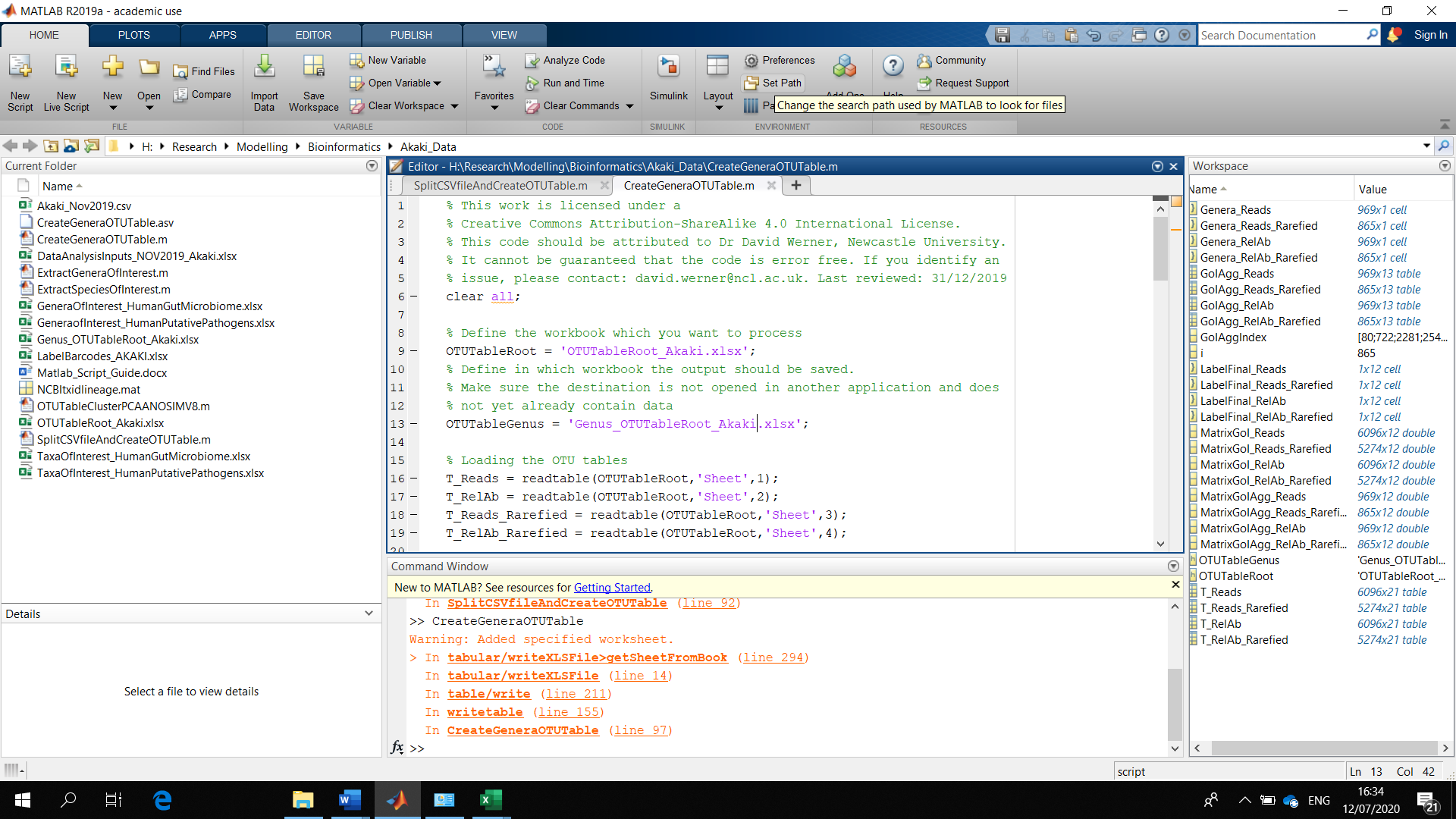
Now you can start interpreting the data by comparing the microbial communities for the different samples. Open the OTUTableClusterPCAANOSIMV8 file. For the ANOSIM to work, you need to add the “FTM” folder (Fathom toolbox) to your Matlab toolbox folder. In my machine the path to the relevant folder is for example

C:\Program Files\MATLAB\R2019a\toolbox

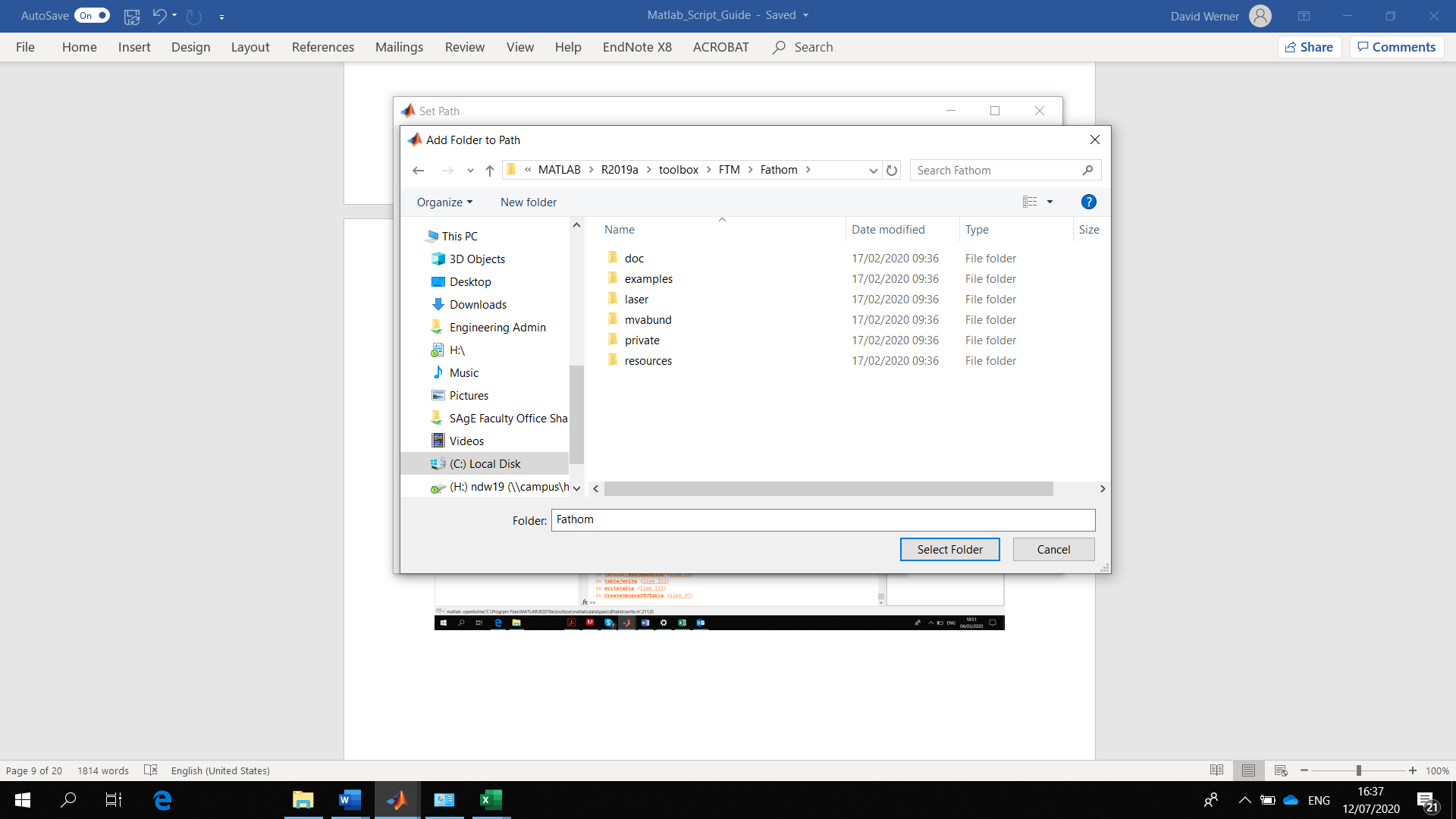
The FTM folder (Fathom Toolbox for Matlab) can be downloaded from this link

<https://www.marine.usf.edu/research/matlab-resources/fathom-toolbox-for-matlab/>

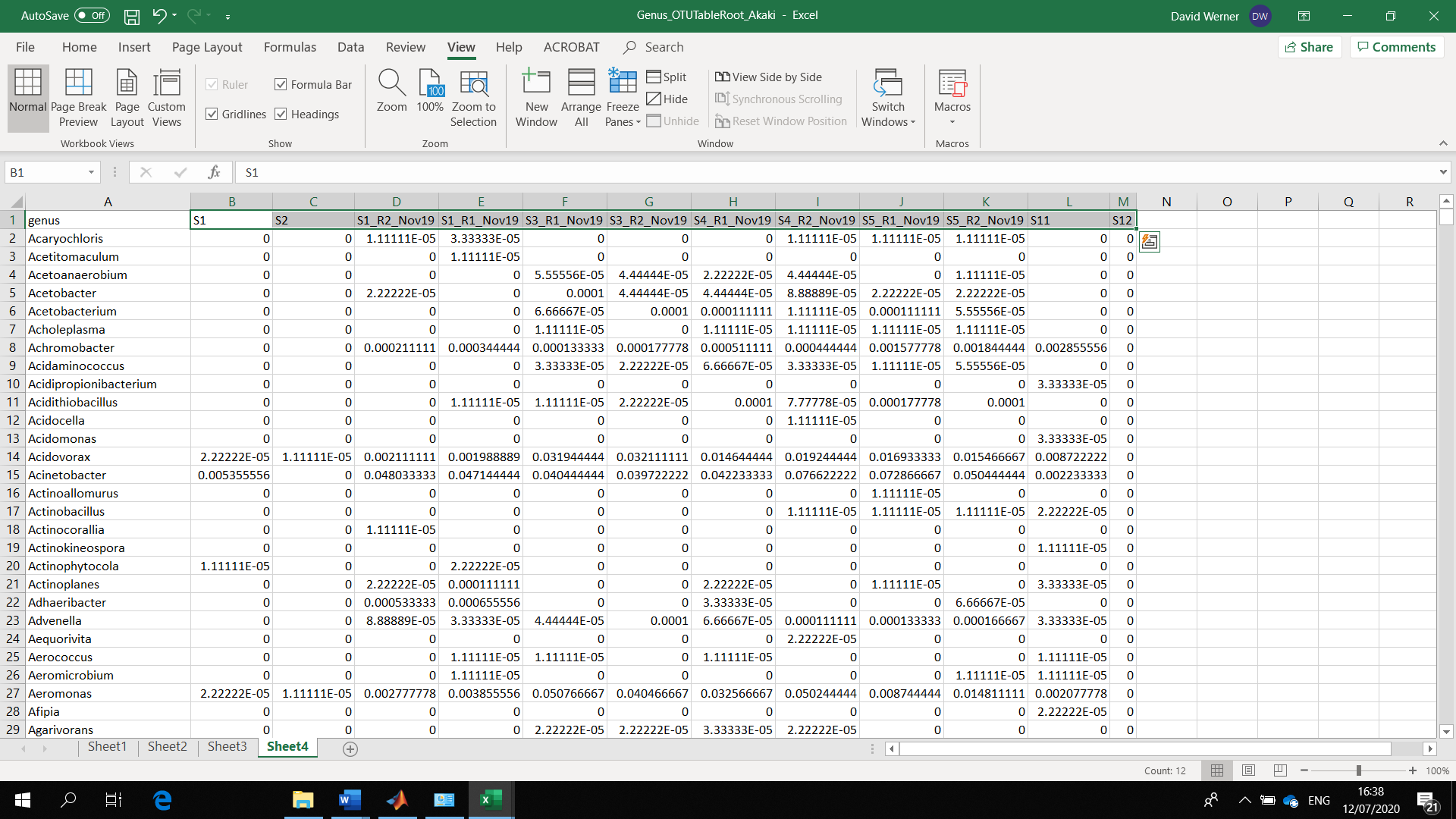
Once your FTM folder is available, you need to set the path to it in the Home tab, using the set path function



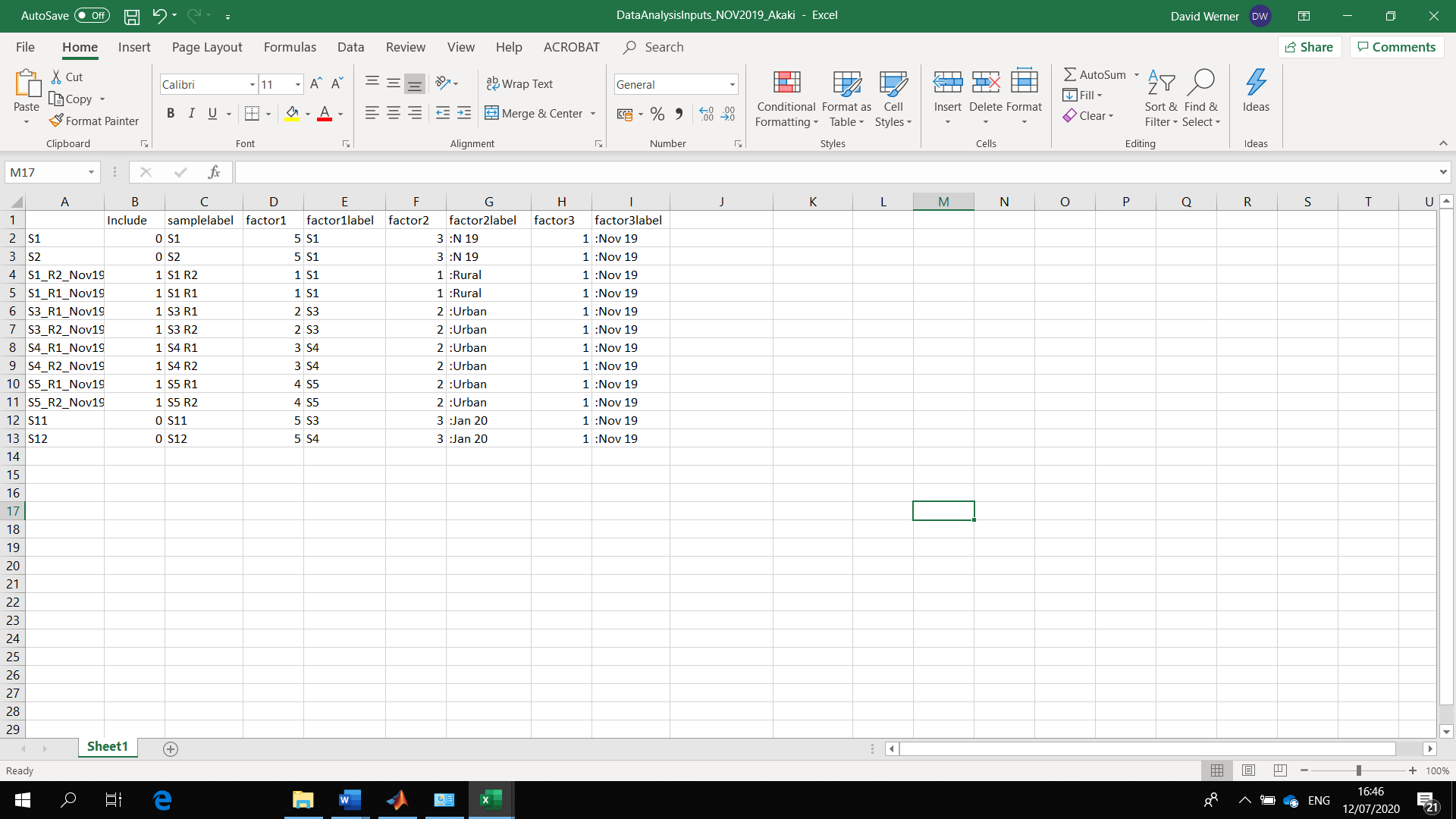
In my labtop the path to the Fathom folder is selected as shown below



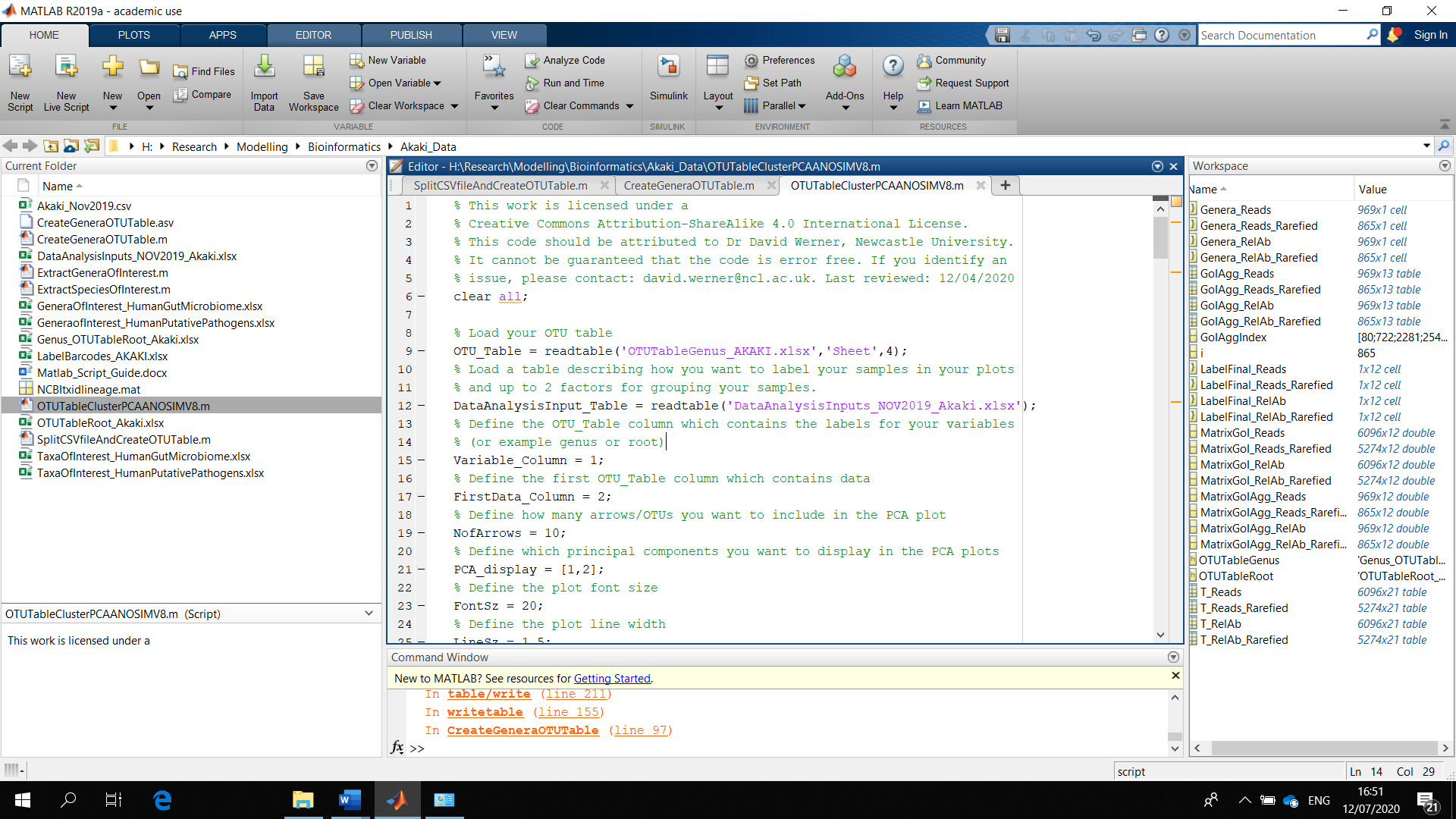
For the data interpretation, you need to explain how your data is structured/can be grouped, and how you want to label it in the plots. In your OTU Table, select the sample labels



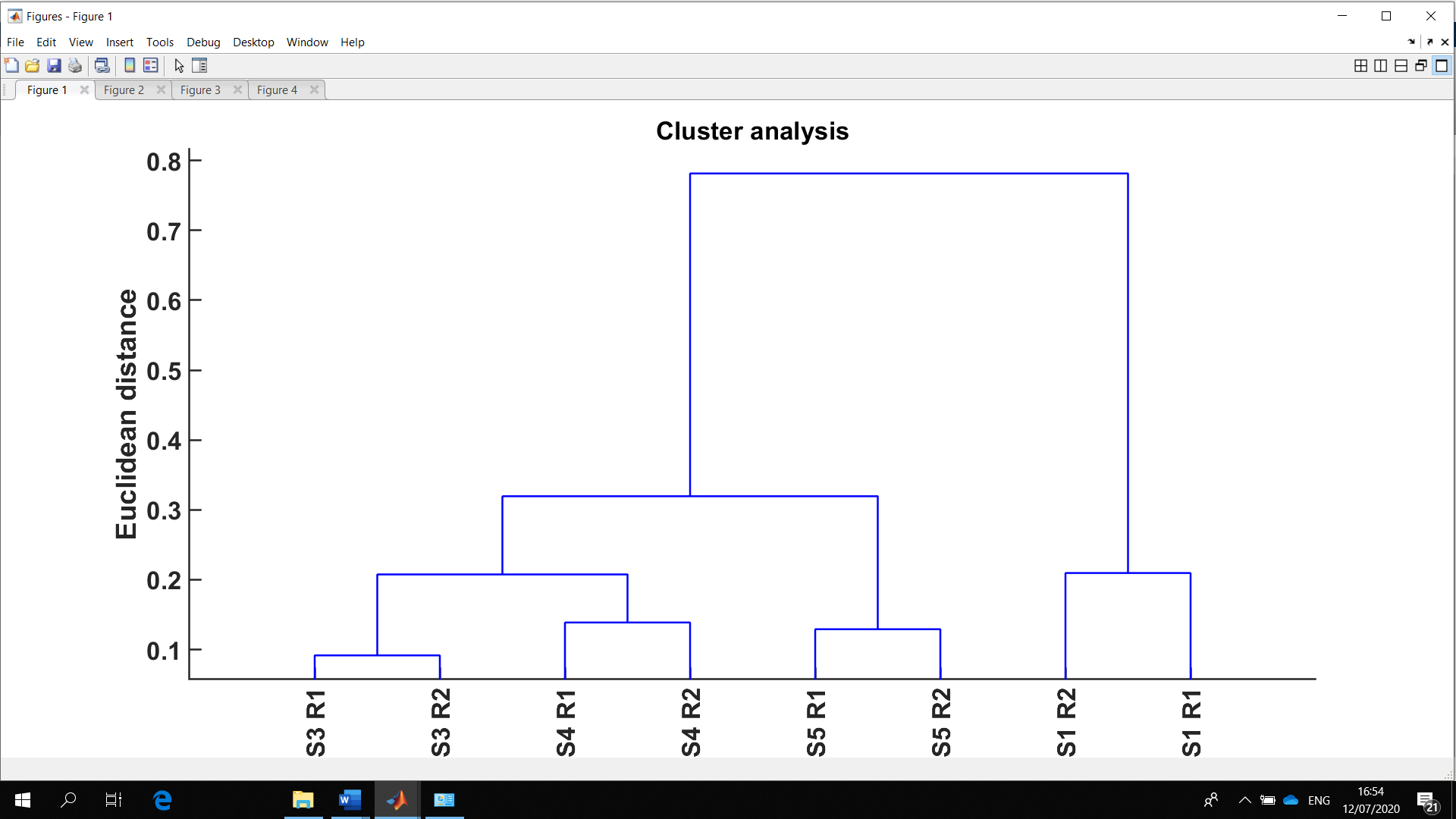
Then use paste special, values, transpose, to copy the labels into the first column of the DataAnalysisInputs spreadsheet (in this example DataAnalysis\_NOV2019\_Akaki). In column B, you need to explain which data you want to include in the analysis of similarities. This can be done by writing “1” to include the sample, or “0” if you don’t want to include it. In column C write the labels you want to use to identify each sample in the plots. There needs to be a label for each sample, even if you don’t want to include it. You can then group your data according to up to three factors. Factor 1 can have up to 15 classifications/groupings, Factor 2 can have up to 4 classifications/groupings, and Factor 3 up to 2 classifications/groupings. In the example, factor 1 is the location where the samples were taken. There are 4 locations. So write 1, 2, 3, 4 into column D (or 5 for no location). There are canal and pond water samples, which were distinguished by writing 1 or 2 into column F (or 3 if unknown). Samples could also be distinguished by writing 1 or 2 into column H, but there was only one sampling event. Labels to distinguish the groupings go into columns E, G, and I respectively. Keep the factor labels short, as these will be combined into a single label, i.e. “S1:Rural” in the example for the first sample. If you have only one factor to group your samples, just write “1” for all samples for factors 2 and 3, etc. Save and close the DataAnalysisInput file when done.



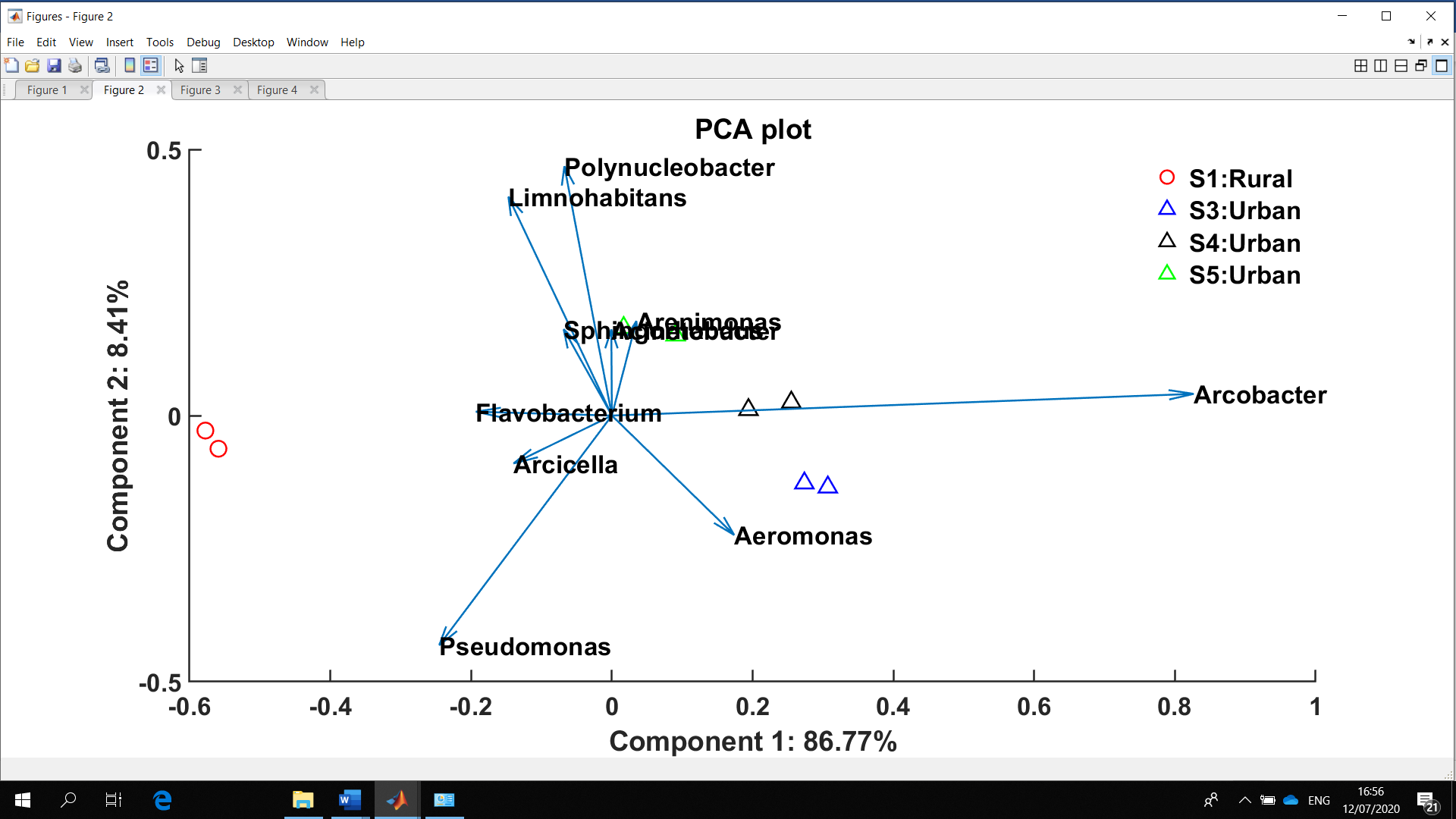
In the OTUTableClusterPCAANOSIMV8 file specify which OTU table you want to interpret (line 9) and which spreadsheet provides the information about the structure of the data and labelling (line 12). You should also state which column from the left in your OTU table contains the variable names (i.e. OTU names, line 15), and which column contains the first sample data (line 17). You can also choose how many variables (i.e. OTUs) to display in your PCA biplot (line 19) and which PCA components to display in the plot (typically that would be the first and second component, line 21). There are some further details of the data analysis and plots which you could potentially modify by making changes in this file. For example, the analysis currently uses rarefied relative abundance data (line 9 opens sheet 4) and square-root transforms that data (line 31). Square root transformed relative abundance data is the same as doing a “Hellinger Transformation” on the original data.



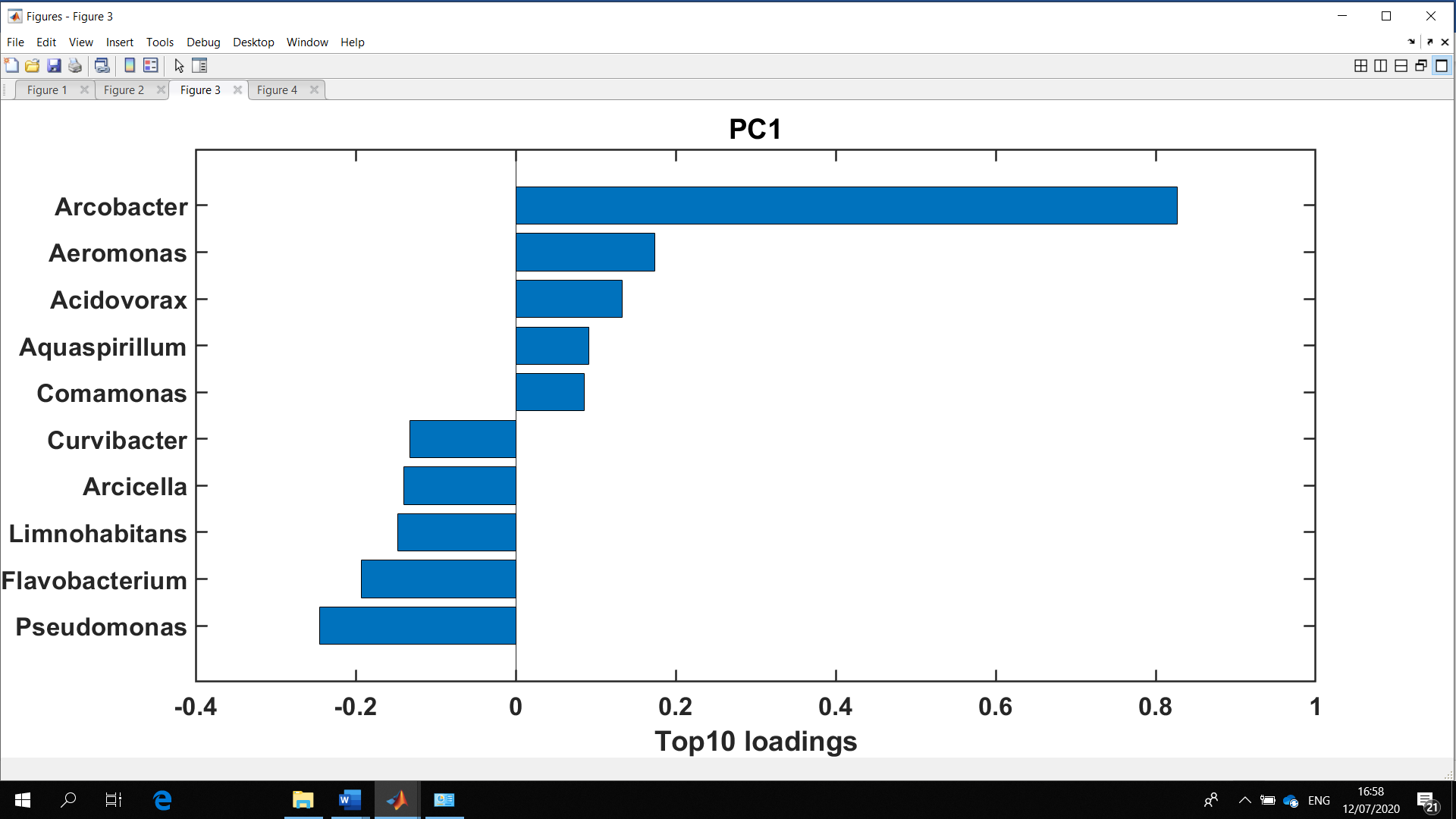
The script performs a cluster analysis, and PCA and tests the significance of the factors in shaping the data using ANOSIM. These evaluations help with an initial exploration of the data. In the example the cluster plot shows a separation of sample S1, and good agreement between replicates.



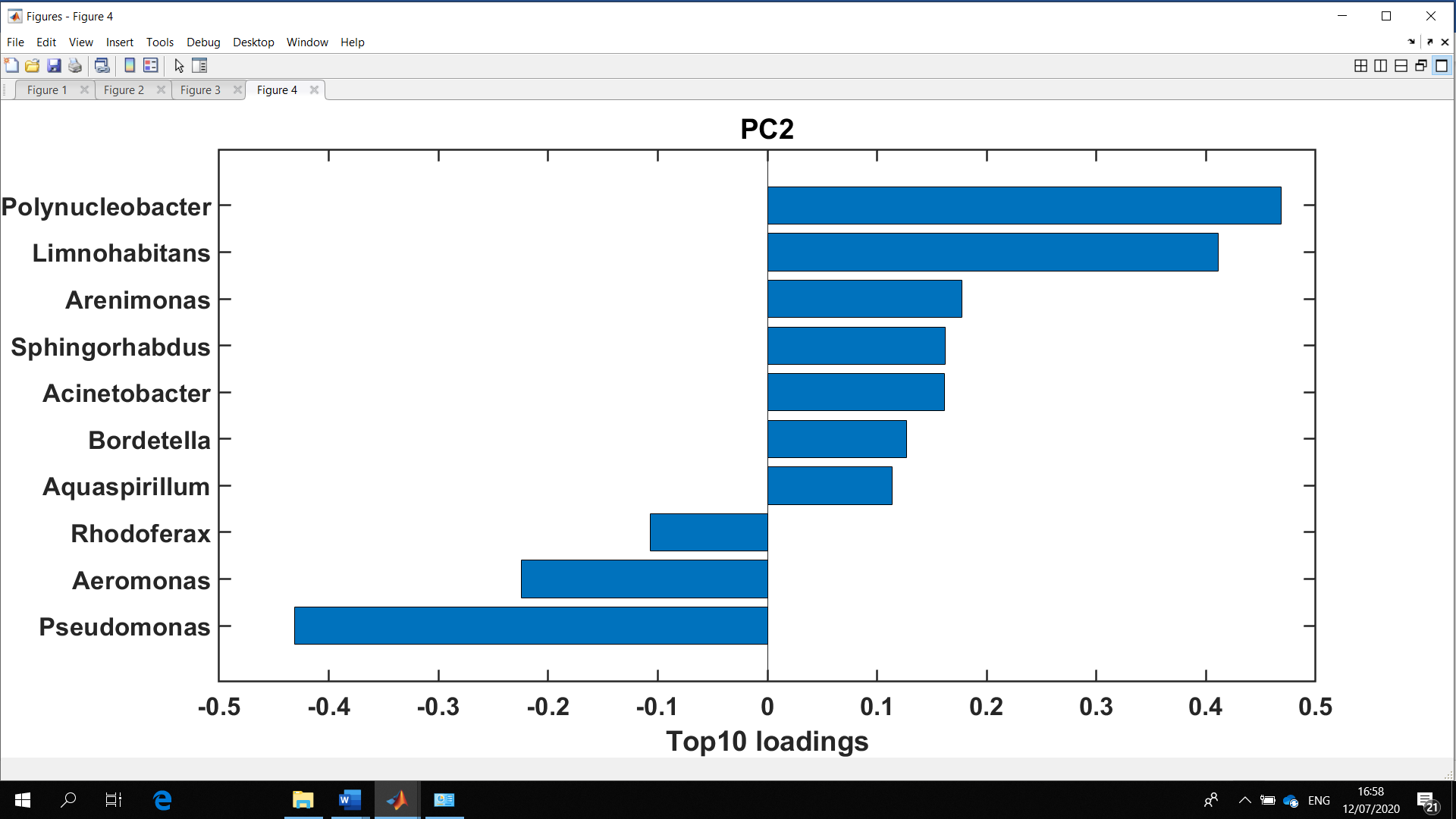
The PCA plot shows a nice separation of sample S1 from the other water samples along principal component 1. The arrows indicate which OTUs are mainly responsible for the separation of samples along these components (loadings).



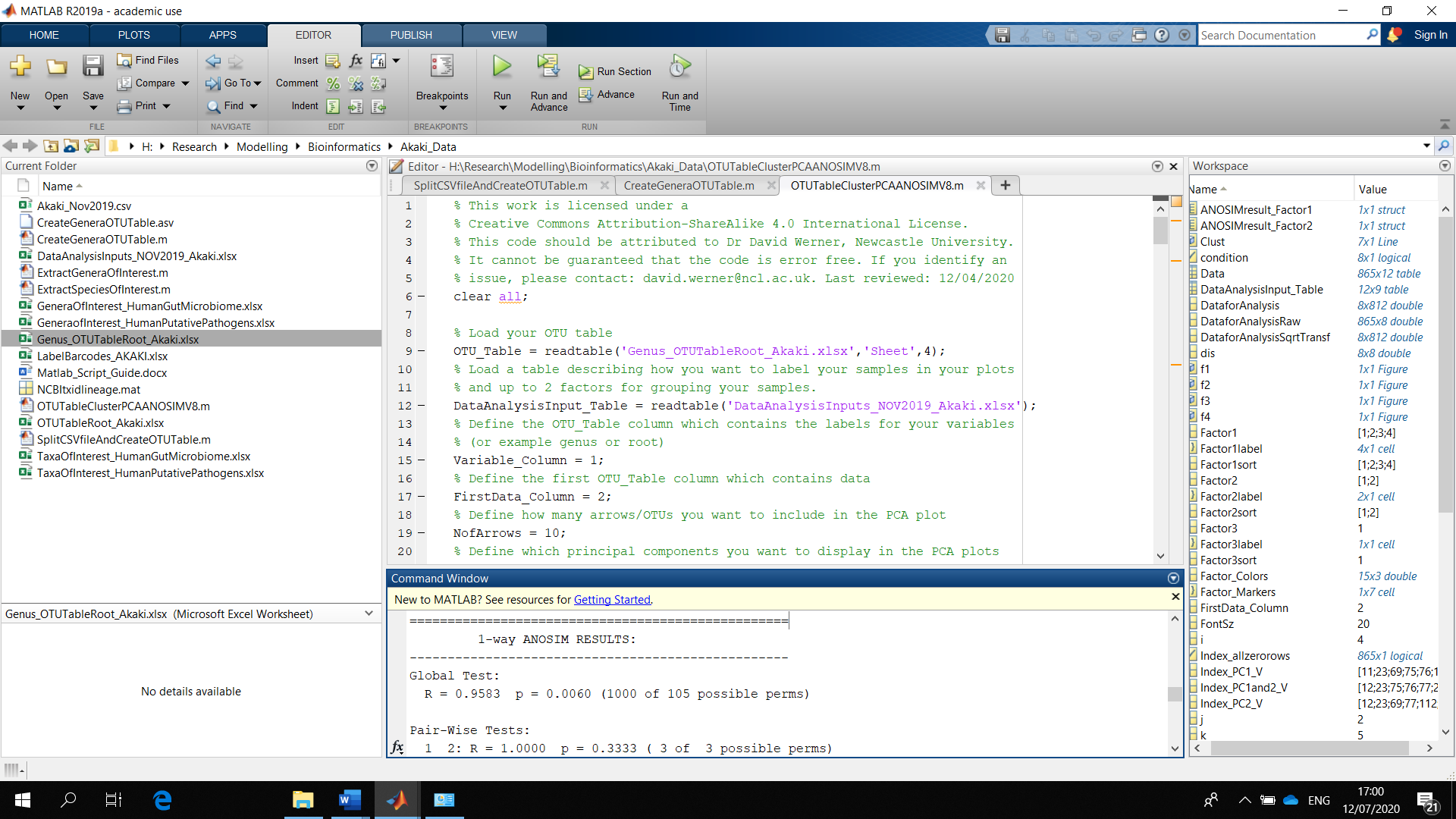
The third plot shows the genera responsible for the top 10 loadings of PC1.



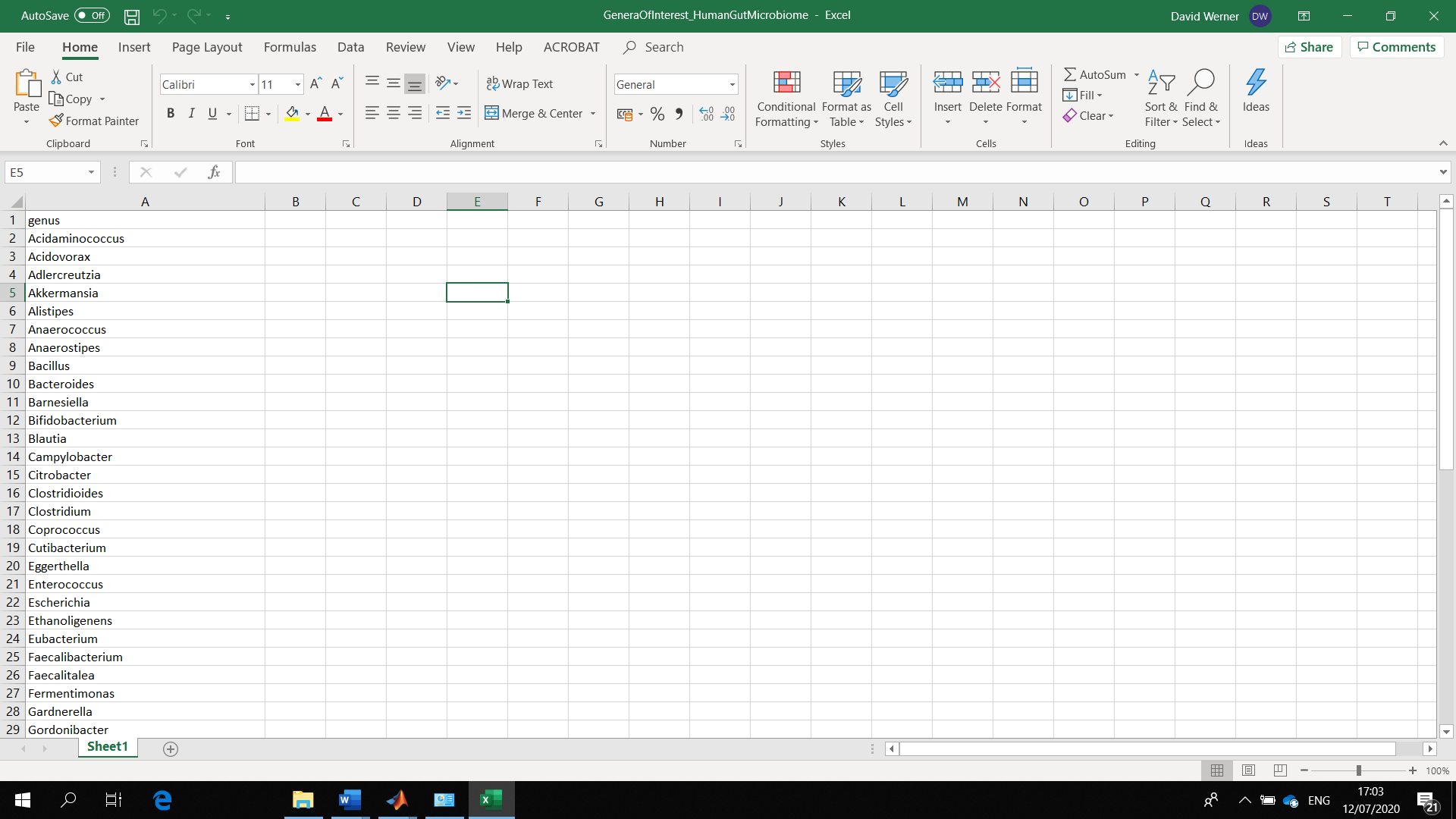
The third plot shows the genera responsible for the top 10 loadings of PC1.



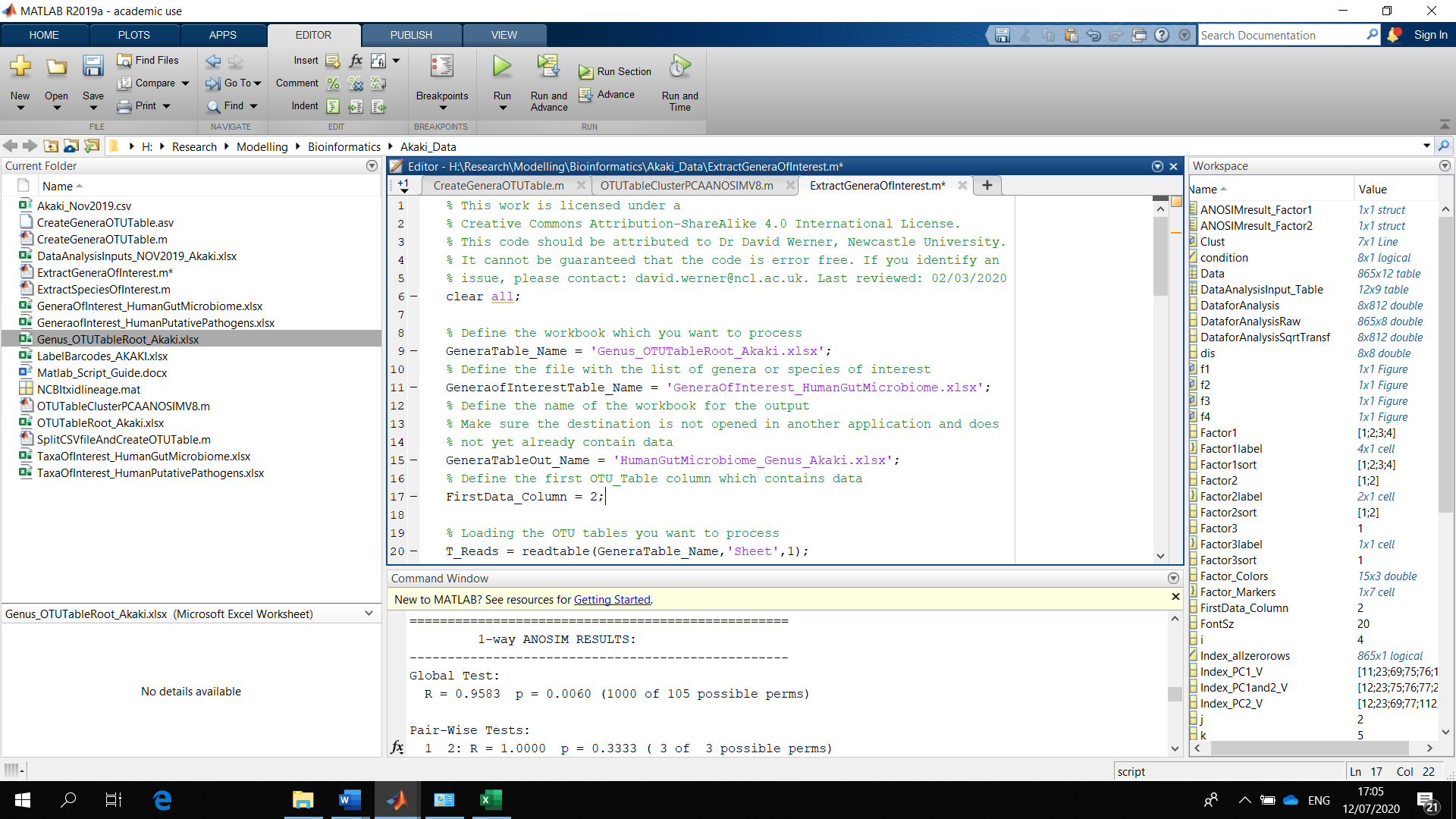
You can see the ANOSIM results in the command window. In the example, the factors have a significant effect on the similarity of samples (p < 0.05), but for the pairwise comparison of sampling locations there is an insufficient number of samples to perform the required permutations for a p < 0.05.



Now that you have investigated the overall OTU table, you can extract data you are particularly interested in, for example genera which are present in the human gut microbiome. For this you need an Excel file in the format shown below. Please note that the spelling of the names needs to be 100% correct, including correct capitalization and for example without an “invisible” empty space at the very end of the “string”, otherwise the matching algorithm won’t work.



Open the ExtractGeneraOfInterest file by double clicking on it. On line 9 explain which OTU table you want to process. On line 11 specify the excel spreadsheet containing the OTUs you are interested in. On line 15 specify the name of the workbook for the output. On line 17 specify which column from the left contains the first sample data set.



The file created is structured in the usual format. There is a similar code to interpret root level OTU tables. Open the ExtractSpeciesOfInterest file by double clicking on it. On line 9 explain which root level OTU table you want to process. On line 11 specify the excel spreadsheet containing the OTUs you are interested in. On lines 18 and 19 specify the name of the workbooks for the output. On line 22 specify which column from the left contains the root level taxa information. On line 24 specify which column from the left contains the first data.

