Using Dot Plots for Comparative Genomic Analyses: Instructions for BLASTn, FASTA file merging, and Gepard

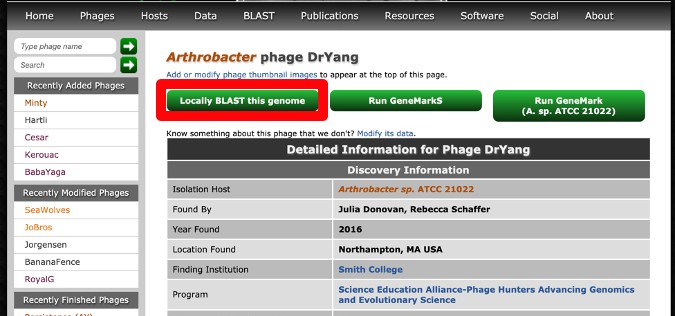
**USING PHAGESDB BLASTN**

Before we begin bioinformatic analyses, it would be very useful to know which phages are most similar to our phages of interest at the nucleotide level. One rapid and simple way to quantitatively describe nucleotide identity is to use the phagesdb BLASTn tool, which is a form of BLAST, the ‘Basic Local Alignment Search Tool.’ To run a phagesdb BLASTn search against your phage of interest, do the following:

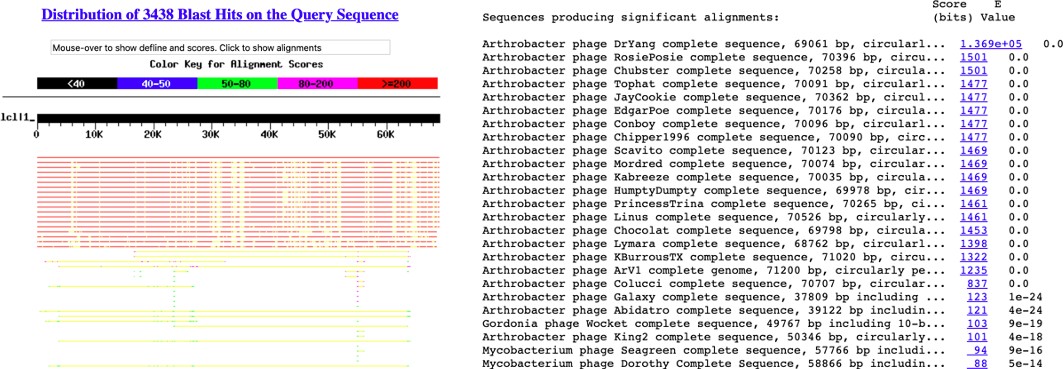
1. Go to [http://www.phagesdb.org](http://www.phagesdb.org/)
2. In the left pane, type in your phage’s name, then click the green arrow to search.



1. Once you are on your phage’s web page, select ‘Locally BLAST this genome’.



1. On the next page, click ‘BLAST’ (in green) to launch the BLASTn search.
2. The BLASTn output page should now load. Notice that there is both a graphic output and a text output for BLASTn. The graphic output includes a colored line for each BLASTn hit, where the color of the line indicates the strength of the alignment score at that region of the genome. The text output includes a list of hits, followed by a more comprehensive description of each hit, as shown below.



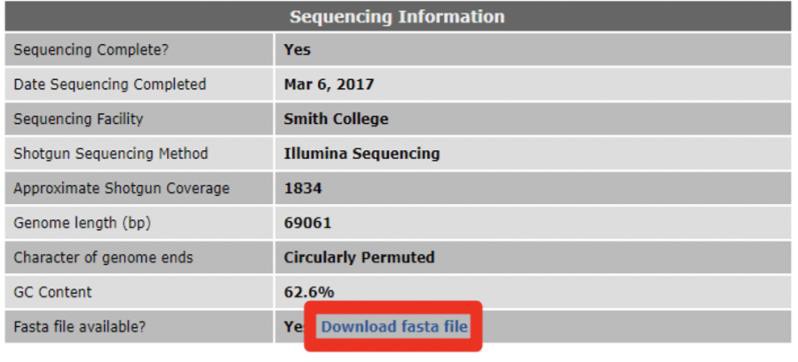
1. When choosing phages to include in your analysis, you should try to include a range of different BLASTn hits. For instance, if you are making a dot plot, you should try to include between 7-12 phages on your first attempt:
   1. 3-5 ‘top’ hits (e-value better than e-10; typically, these phages infect the same host and belong to the same cluster as the phage of interest).
   2. 3-5 ‘okay’ hits (e-value between e-3 and e-10; perhaps infecting different hosts or belonging to different clusters).
   3. 1-2 ‘poor’ hits (e-value worse than e-3).
2. Once you have selected the phages you would like to include in your comparison, you should document the following in your notebook:
   1. Screenshot of graphic **AND** text outputs
   2. Names of phages being compared
   3. Scores of phages being compared
   4. E-values of phages being compared
3. The next step is to download the FASTA files of your phages from phagesdb.org. This is described in the next section.

**DOWNLOADING FASTA FILES (NT ONLY) FROM pHAGESdb**

1. Type the phage name into the phagesdb search bar and click on the green forward arrow.



1. On the phage’s web page, scroll down to the section titled ‘Sequencing Information,’ then click on ‘Download FASTA file’.



1. Once the file has been downloaded, locate it in your Downloads folder and move it to a folder on your computer. There, you can rename the file to include the phage cluster and change the file extension from ‘.fasta’ to ‘.txt’ (eg. ‘Tanis.fasta’ should be changed to ‘DJ\_Tanis.txt’).
2. Once the file has been appropriately named, you can double-click it to open (if prompted to select an application to open the file in, select either ‘Text Editor’ for MacOSx, or ‘Notepad’ for Windows).
   1. On the first line, you will see some identifying information for the FASTA sequence. You should remove most of this information to ‘clean up’ the file, which makes it easier to use later on (eg. when trying to find phages on the graphic output of your dot plot). Changing this first line to include the phage cluster, name, and host only will be sufficient. (eg. change the first line of the file to read ‘> Tanis\_DJ\_Gordonia\_terrae).
   2. Note that the ‘>’ symbol is required for the bioinformatics software to recognize the file and should not be removed.

**DOWNLOADING GENBANK AND WHOLE GENOME FASTA FILES (NT OR**

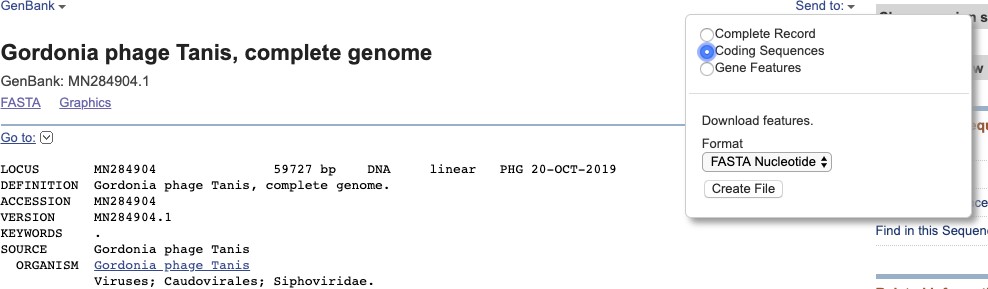
**AA) FROM NCBI**

The GenBank file and full genome nucleotide and amino acid sequences of a manually annotated phage can be obtained using the NCBI website (<https://www.ncbi.nlm.nih.gov/nuccore/>).

The following protocol is initiated from the phagesdb site, but you can also try searching for your phage directly on NCBI GenBank. If searching directly from GenBank, you should be sure to search for the phage by host and phage name (eg. ‘Gordonia phage Tanis’).

1. Search for your phage on phagesdb.
2. On your phage’s page, ensure that the genome has been published to GenBank (this will be under the ‘Publication Info’ section, see below). If the final genome annotation has been uploaded to GenBank, you can click on the blue hyperlinked accession number to go to that phage’s NCBI ‘Nucleotide’ page.



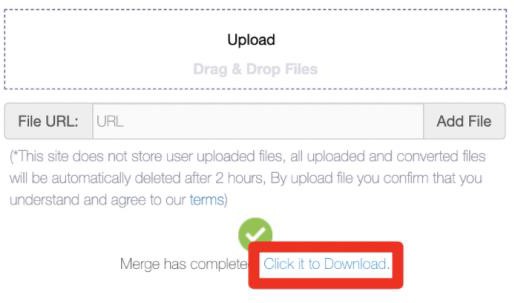
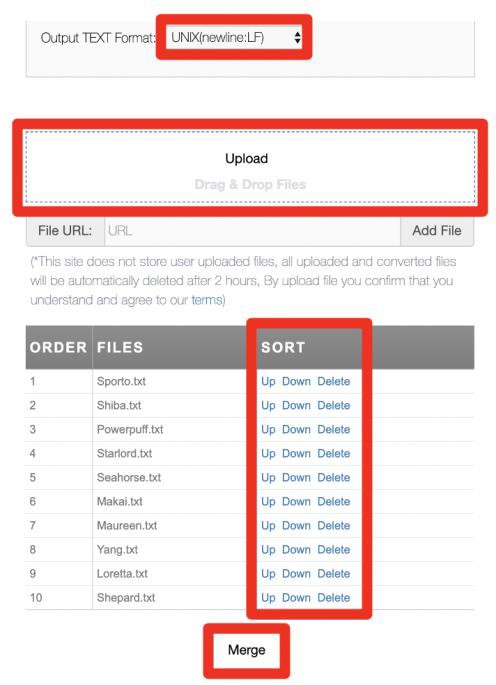
1. On the NCBI GenBank site, confirm that you’ve accessed the correct record by double- checking the phage name at the top of the page. When you are ready to download your file, select the following options from the ‘Send to’ dropdown menu:
   1. Whole genome nucleotide FASTA: Send to > Coding Sequences > FASTA Nucleotide > Create File
   2. Whole genome amino acid FASTA: Send to > Coding Sequences > FASTA Protein > Create File
   3. Whole genome GenBank file: Send To > Complete Record > File > GenBank > Create File

**MERGING FASTA FILES**

Gepard dot plot software requires a single input file for each axis, so in order to compare multiple phages, we must first merge our files to contain multiple full-length genome sequences. Once we have generated our dot plot, we may also wish to reorganize the order of the phages along each axis, in order to ensure that the most closely related phages are adjacent to one another. To merge multiple FASTA files, complete the following:

1. Go to <https://www.ofoct.com/merge-text-files-online>
   1. Select ‘UNIX’ for output format.
   2. Drag and drop your files into the ‘Upload’ box.
   3. Rearrange as needed, then click ‘Merge’; preliminarily, you should ensure that all of the phages that belong to the same cluster are grouped together on the file list.
2. Once your files have been merged:
   1. Download the merged file by clicking on the blue hyperlink that says ‘Click it to Download’.
   2. Rename the file descriptively, then document file name and phage order in your notebook for future reference (eg. ‘103AL-S20\_AZ-EH-EB-Singletons.txt’).

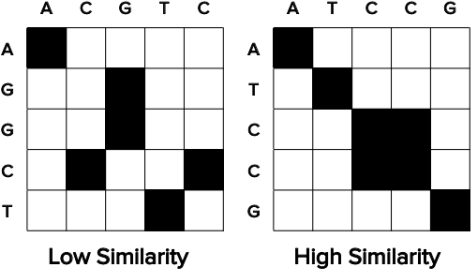
***Hint:*** *keep this window open in case you need to rearrange your files later. This happens often when generating dot plots, as we want to make sure that the most closely related phages are next to each other on the graphic output, and it may take some trial and error to achieve this.*



**GEPARD DOT PLOT SOFTWARE**

Dot plots are a graphical representation of the similarity between two sequences (either nucleotide or amino acid) and can be used to quickly analyze the whole genome relationships between large groups of phages.

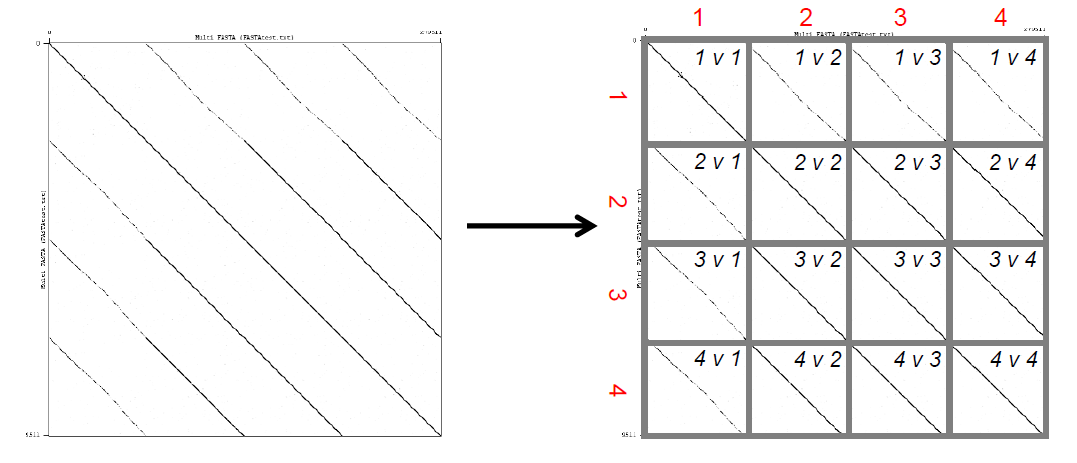
In a dot plot comparison, sequences are not compared to one another altogether.

Instead, short chunks of nucleotides/amino acids, called ‘words,’ are compared. To generate a dot plot, the user will specify a ‘word length,’ which describes the stringency of the comparisons between the sequences. The ‘word length’ is, in essence, the number of nucleotides (or amino acids) which must be an exact match in order to place a dot on the plot. In the example here, the word length/size has been set to 1, meaning that each time any one nucleotide is found along both sequences, a dot is placed at that coordinate of the 2D plane. Note how, when the sequences are ***dissimilar***, there appears to be weak, scattered signals. However, when the sequences are ***similar***, there appears to be a thick, strong diagonal signal on the graphical output, indicating that the sequence has been well conserved.

The same goes for larger sequences and larger word sizes. The figure below includes a nucleotide comparison of four phage genomes, using a word size of 15. This means that, every time there is an exact match of 15 nucleotides on both sequences, a dot is placed on the graph. Also, in this case, we have merged the FASTA files containing the whole genome sequences of the phages being compared, meaning that each axis of the dot plot contains four full-length genomes that have been ‘glued’ together. By making the horizontal and vertical input sequences the same, we can generate a multi-FASTA dot plot that allows for informative comparisons between many phages at the same time. You can think of each ‘sector’ where genomes intersect as ‘mini-dot plots’ for just those two phages. Note that, where self-to-self comparisons are made (eg. 1 v 1 below), the signal is always strong, as the sequence being compared is identical at these regions.

1

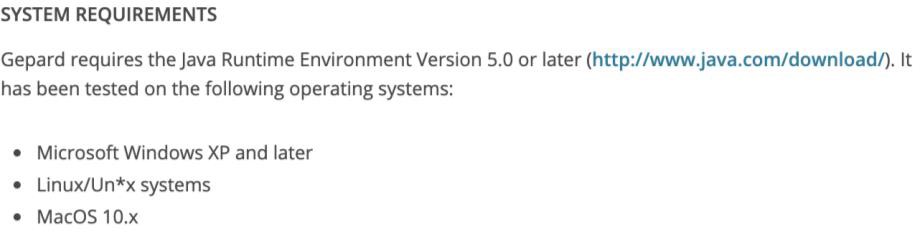
2



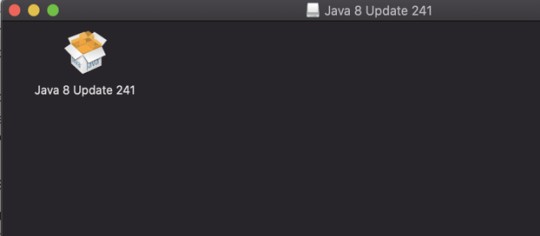
# Setup for MacOS 10.x

1. Go to [http://cube.univie.ac.at/Gepard](http://cube.univie.ac.at/gepard) and scroll down until you reach the ‘System Requirements’ section seen below.

*You must have one of the following operating systems in order to use Gepard. You will also need to download Java in order to use the software.*



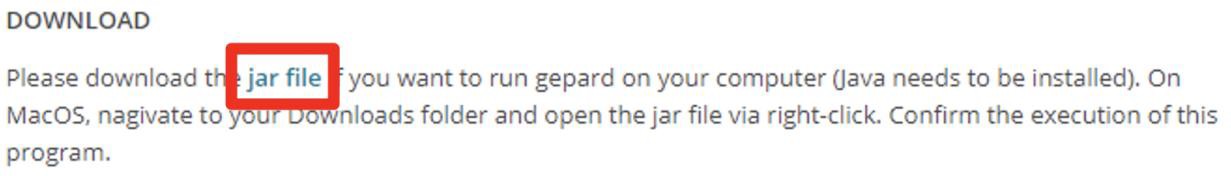
1. On the Gepard site, click the Java download hyperlink. You will be prompted to download Java.
2. Double click the icon in the popup window to begin the download.



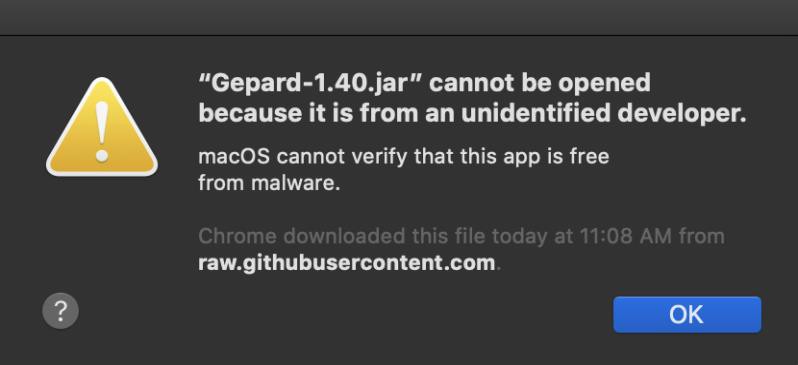
1. Enter your computer account’s password if prompted by the system to do so.
2. Click ‘Install’ in the Java window that opens.



1. After installing Java, return to the Gepard website and scroll down to the ‘Download’ section.
2. Click the jar file hyperlink to download Gepard.



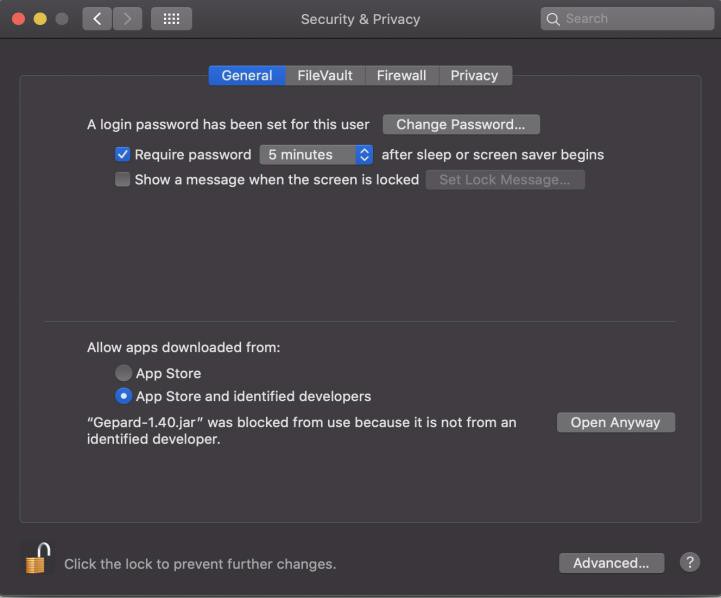
1. The following window will likely pop up. Click ‘OK’.



1. Open ‘System Preferences’ and click on ‘Security & Privacy’.



1. In ‘Security & Privacy’ go to the ‘General’ tab and click ‘Open Anyway’.

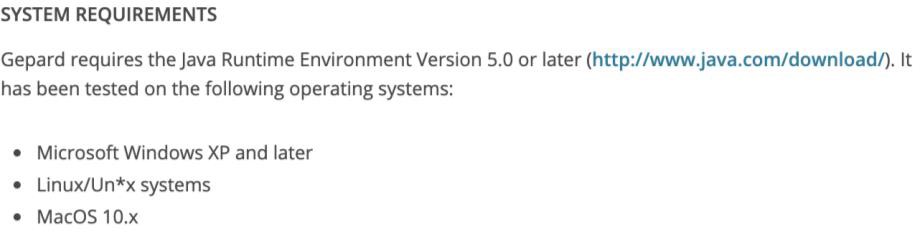


1. Gepard will download. You are now ready to start creating dot plots!

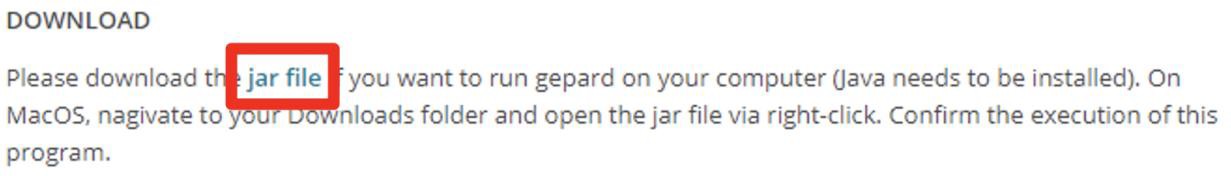
# Setup for Windows 7 and later

1. Go to [http://cube.univie.ac.at/Gepard](http://cube.univie.ac.at/gepard) and scroll down until you reach the ‘System Requirements’ section seen below.

*You must have one of the following operating systems in order to use Gepard. You will also need to download Java in order to use the software.*



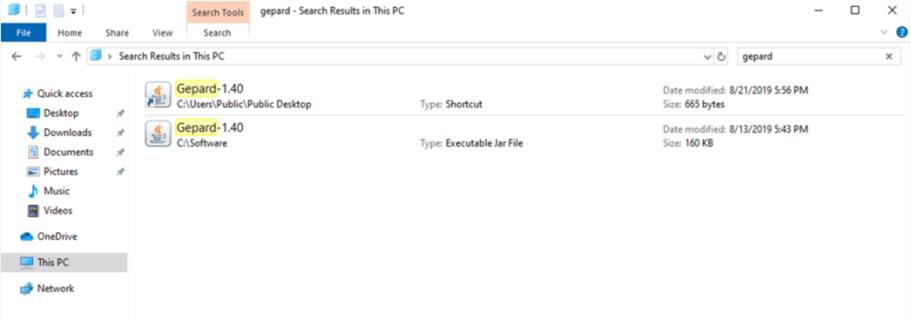
1. On the Gepard site, click the Java download hyperlink. You will be prompted to download Java.
2. Double click the icon in the popup window to begin the download.



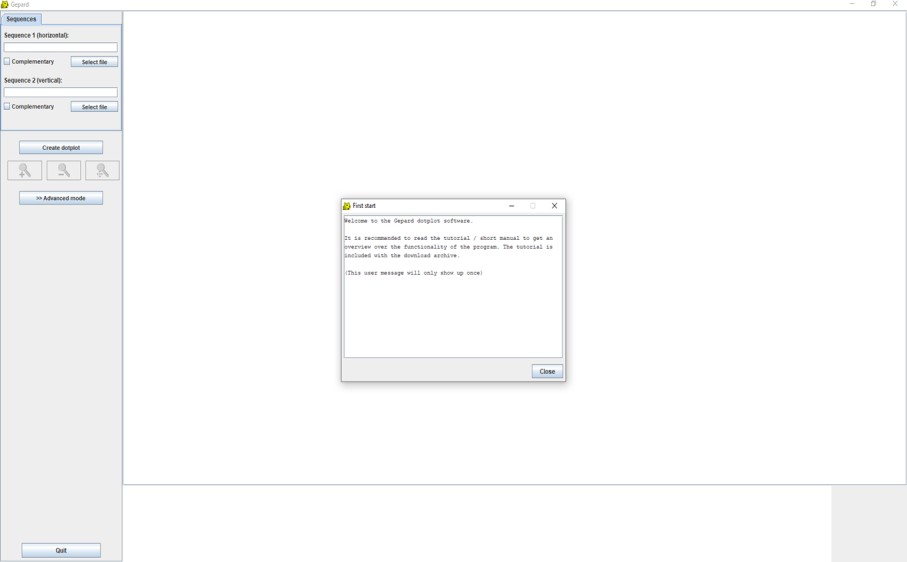
1. Gepard will download. You are now ready to start creating dot plots!

# Gepard Walkthrough

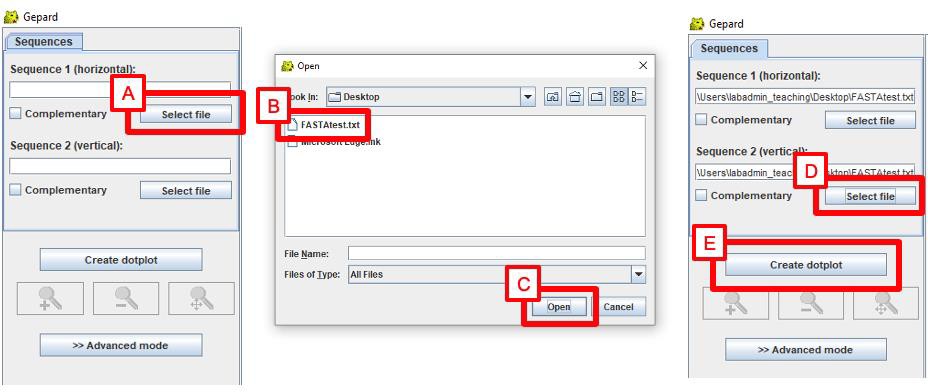
1. To begin, open the Gepard dot plot software. Start by searching your computer for the Gepard jar file, then double clicking to open.
   1. MacOSx users: Open Finder, search for ‘gepard-1.40’ or ‘gepard’ (depending on version; the file will have a .jar extension).
   2. Windows users: Open Files, search for ‘gepard-1.40’ or ‘gepard’ (depending on version; the file will have a .jar extension, see below).



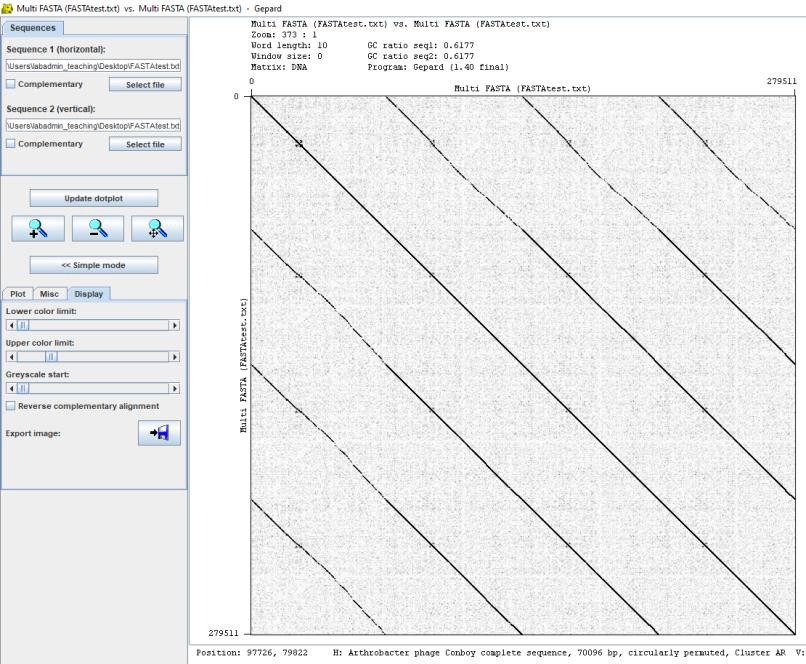
1. The Gepard window will open (see below).



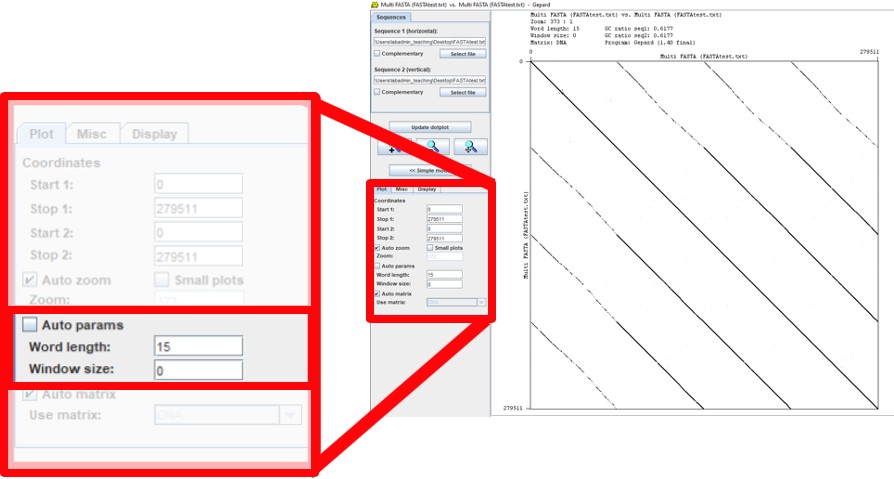
1. Complete the following:
   1. In the left pane, click ‘Select file’ to upload an input sequence for Sequence 1 (horizontal).
   2. Select the input file containing the sequence for which you would like to generate a dot plot (and be sure to ***document the name of this file in your notebook***).
   3. Select ‘Open’ to proceed.
   4. Next, click ‘Select file’ to upload an input sequence for Sequence 2 (vertical). Select the ***same*** input file as you did for Sequence 1. This will ensure that both axes of the dot plot are the same.
   5. Select ‘Create dot plot’ to generate your dot plot using automatic parameters (word size of 10).

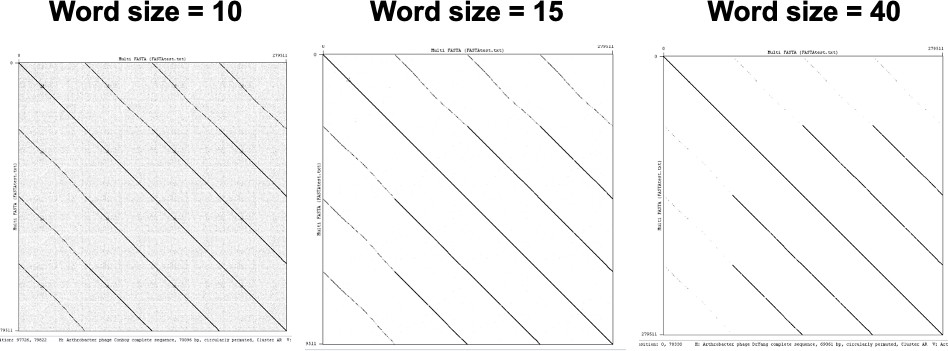


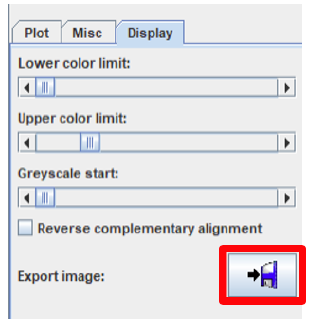
1. A dot plot should now be generated in the main frame of the Gepard window (see example output below).



1. You may wish to change the word size of your dot plot. You can do so by unchecking ‘Auto params’ under the ‘Plot’ tab of the left pane, then entering a new word size in the ‘Word length’ window. Click ‘Update dot plot’ to update the graphic output.
   1. Consider ***increasing*** the word size if there is too much background noise on your dot plot, or if there are too many signals on your plot. This may be an indication that the word size is too low, and that there are too many similarities between the sequences at that stringency.
   2. Consider ***decreasing*** the word size if you are unable to discern any signals on your dot plot. This may be an indication that the word size is too high, and that the parameter is too stringent to capture subtle similarities.







1. There are many other settings you can change to fine tune your dot plot. Check out the Term Glossary below to investigate the rest of the software’s tools.
2. Once you’re satisfied with your dot plot, export the graphic output as a ‘.png’ file, then upload the image file to your notebook.

# Gepard Term Glossary



1. **Select file:** Select this to upload the fasta/.txt file for sequence 1, which will be displayed along the horizontal axis.
2. **Select file:** Select this to upload the fasta/.txt input file for sequence 2, which will be displayed along the vertical axis (for self-to-self comparisons of a multi-FASTA file, you should upload the same file as you did for sequence 1).
3. **Create/update dot plot:** Once you have selected all of the appropriate settings from the ‘Plot,’ ‘Misc,’ and ‘Display’ tabs, you can select this option to generate your dot plot. Any changes made to the dot plot settings will only be reflected on the graphic output *after* this button has been reselected.
4. **Zoom in:** Select this option to ‘Zoom In’ on your dot plot’s graphic output. If you have highlighted a region on your graphic output, selecting this will automatically rescale the highlighted region to fill the screen.
5. **Zoom out:** Select this option to ‘Zoom Out’ on your dot plot’s graphic output.
6. **Reset zoom:** Select this option to return to the full view of your dot plot.
7. **Toggle simple/advanced mode:** Select to toggle between ‘Simple’ mode (no additional settings) and ‘Advanced’ mode (displays additional settings for ‘Plot,’ ‘Misc,’ and ‘Display’).
8. **Plot tab:** Select this tab to configure various settings, such as the coordinates of the graphic output being displayed, zooming in/out, word length, and matrix type.
9. **Misc. tab:** Select this tab to configure various settings, such as the management of suffix arrays and toggling of local click options for your graphical output.
10. **Display tab:** Select this tab to configure various settings, such as the upper/lower color limits and greyscale settings, as well as to export your graphic output as an image file.
11. **Coordinates:** Allows you to change the start and stop coordinates for the region of the graphic output being displayed. Start & Stop 1 refer to the horizontal axis (sequence 1), while Start & Stop 2 refer to the vertical axis (sequence 2).
12. **Auto zoom:** This option is selected by default. If you would like to change the zoom magnitude, you may deselect this option and enter a value in the adjacent box.
13. **Small plots:** This option is deselected by default. Selecting this option will decrease the size of the graphic output without adjusting the zoom magnitude.
14. **Auto parameters:** By default, the word size for Gepard dot plots is set to 10 base pairs/amino acids. To change this word size, you must deselect ‘Auto Params.’
15. **Word size:** To adjust the word size for your dot plot, ensure that ‘Auto Params’ is not selected, then enter your new word size into the ‘word size’ box. *Note: to update your graphic output to reflect the new word size, you will need to select ‘Update Dot plot’.*
16. **Auto matrix:** By default, Gepard will assume that you would like to process a DNA matrix (i.e. that your input files contain FASTA nucleotide sequences). The DNA matrix setting will also allow you to make a dot plot with amino acid sequences.
17. **Use matrix:** You may change this setting to use an alternative matrix for processing your inputs. Please note, however, that the DNA matrix setting is sufficient for comparing FASTA files containing either nucleotide or amino acid sequences.
18. **Save suffix arrays:** Suffix arrays are generated by Gepard when the sequence matrix is first processed. By selecting ‘save suffix arrays,’ Gepard will automatically save the suffix array used to generate your dot plot, thus reducing the computation time required to generate the graphic output when you recreate/update the graphic output using the same input files. *Note: computation times tend to be short for small datasets. Thus, it may not be necessary to save suffix arrays for multi- FASTA input files containing fewer than phage 100 genomes.*
19. **Delete suffix array files:** You may select this option to clear any saved suffix array files.
20. **Local dot plot click action:** You may toggle this setting to change the outcome of a local click on your dot plot’s graphic output.
    1. Selecting ‘**...show alignment**’ will display in the bottom pane the local alignment of your sequences at the coordinates you selected on the graphic output.
    2. Selecting ‘**...export window to fasta**’ will export a copy of the sequence 1 and sequence 2 input files to a destination folder of your choosing.
    3. Selecting ‘**...do nothing**’ will not result in any outcome.
21. **Lower color limit:** Adjusting this scale bar will increase/decrease the contrast of any dots which are distant from adjacent dots (i.e, it will make weak signals in the graphical output darker/lighter).
22. **Upper color limit:** Adjusting this scale bar will increase/decrease the contrast of any dots which are proximal to adjacent dots (i.e. it will make strong signals in the graphical output darker/lighter).
23. **Greyscale start:** Adjusting this scale bar will darken the lightest possible color on the graphical output of your dot plot (i.e. change the background from white to grey/black). The default greyscale is set to the minimum and should typically not be changed.
24. **Export image:** Select this option to export the graphical output of your dot plot to an image file. You may change the destination folder using the drop-down at the top of the pop-out window. Images can be saved as a .jpeg, .png, or .bitmap. *Note: exporting your file using this option may result in a low-resolution image. You may want to take a screenshot of the graphic output directly to save a high-quality image for your dot plot.*