**Making trees from PhyloSNP pipeline**

PhyloSNP is a powerful tool that allows for generation of trees from human, bacterial, or viral genomes. This tutorial will guide you through the steps of installing the PhyloSNP shrunk-genomes tool, the preferred tree building software FastTree, running analysis and creating publication quality images of the results.

**INSTALLATION OF PROGRAMS**

**Install PhyloSNP**

Log into your hive shell via any terminal program. From the command line, type wget “https://hive.biochemistry.gwu.edu/static/phylosnp/PhyloSNP\_Unix.zip” and hit ENTER. This will download the zipped PhyloSNP package to your home directory. Next, unzip the package by inputting unzip PhyloSNP\_Unix.zip. This will create a folder named PhyloSNP\_Unix. Since this will be the directory will be where all analyses will run, it is recommended that you change this directory name by entering mv PhyloSNP\_Unix name\_of\_new\_dir. Use cd name\_of\_directory to change directories and enter ls to double check that the folder contains the files genome\_example.csv, inc, phylosnp.pl, Readme.txt, shrunk-genomes.pl. Next we will download and install FastTree and the hg19 reference genome for human analysis. To do this, while in the same directory, download the hg19.2bit file from the ucsc website by entering wget --timestamping “ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/hg19.2bit”. Now, since this file will be in a binary format that needs to be converted via a program described in the next section

**Install FastTree and convert hg19.2bit to fasta**

Change to your local bin folder by using the command cd ../bin. From here, we are going to download FastTree. To do so, type wget --timestamping “http://meta.microbesonline.org/fasttree/FastTree.c”. Compile FastTree using gcc –Wall –O3 –finline-functions –funroll-loops –o FastTree –lm FastTree.c. Now that FastTree has been compiled, we will download twoBittoFa by using the command wget --timestamping “http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86\_64/twoBitToFa”. Next, input chmod 755 twoBitToFa to make the file an executable. Before we exit the bin folder, let’s make sure that our path is set correctly by entering echo $PATH. Look at the output and make sure that you see /home/your\_usr\_name/bin listed. If it is not, type echo $PATH:/home/your\_usr\_name/bin to add your bin folder to the path allowing for universal file execution. Now there is one last step before we can start analysis. Use cd ../name\_of\_new\_dir to switch back to the PhyloSNP directory. By typing twoBitToFa hg19.2bit hg19.fa followed by rm hg19.2bit after completion of the previous command, you will unpack the hg19 genome and delete the old 2bit file.

**DATA ANALYSIS**

**Add Files to directory**

For this step I use FileZilla to transfer my folder containing my data to my PhyloSNP folder via a drag and drop method. Once I have the files into my PhyloSNP folder, cd to change to the files directory and I use more (hit enter to see new lines) to see the names the names of the columns I need (either in .vcf or .csv format). The columns I need are the chromosome (reference) column, a position column, and a change column. Once I have this information I back out of the more by hitting q and then typing cd .. to get back to the parent directory where shrunk-genomes.pl is located.

**Shrunk-Genomes Data Analysis**

To start an analysis, use ssh hive3 to move to Hive3 to perform analysis and navigate to your PhyloSNP directory. Enter perl shrunk-genomes.pl dir\_of\_your\_files --position-col=POS(change) --change-col=Change(change) --reference-col=CHROM(chrom) --reference-file=hg19.fa --position-delta=0(change) > output\_file.fasta &. Note where it says (change) change to the appropriate column name you found above, and the number of flanking positions desired in the position-delta. Also note that the & command allows the user to run shrunk-genomes.pl in the background after logging off. If this is your first time using the data set it is recommended to add nohup at the beginning of the above command to output any errors to the output file, however, this will require downloading the output file, removing the shrinking genomes... line and re-uploading the fasta file. This can take up to several hours to execute depending on the size and amount of the files, so it is recommended to run this overnight.

**Using FastTree and FigTree**

After the shrunk-genomes is finished, it is time to run FastTree, which will take anywhere from a couple of minutes to an hour or more, depending on size and options chosen. For this pipeline, we will just run the neighbor-joining algorithm in the program. In your data directory, enter FastTree –nt –nome –noml < output\_file.fasta > treefile.tre &. Once the file has been generated, download to your local machine and open in a text editor. The file will look like this:

|  |
| --- |
| Newick Format FastTree File |
|  |

Now, this file can easily be loaded into a tree viewing program, but I prefer to convert the file into Nexus format, which consists of a table of sample ids and a number corresponding to them followed by a newick format output. See <http://en.wikipedia.org/wiki/Nexus_file> for more info. Once I have renamed my samples and given them ID values, it will look something like the image below.

|  |
| --- |
| Completed Nexus format |
|  |

**Tree Building**

Now, it is time to make some trees. I use FigTree (<http://tree.bio.ed.ac.uk/software/figtree/>) for my trees. Upon first opening the tree, you will be prompted for a label for your support values. Just hit enter and open the trees. This will be the first look at your tree.

|  |
| --- |
|  |

The next step is to increase the thickness of the lines, remove the scale bars, add support values, and line up sample names. Click Appearance and increase line width to 4. Next, select Trees, check Transform branches, and select proportional. Check branch labels, select label(or what you entered), and click Font. Here, select Arial, bold, 12 for the font selection. Finally, uncheck Scale Bar. Now your completed tree

should look like this:

|  |
| --- |
|  |

This tree looks fine for general analysis but pretty boring for publications. In order to make the tree more appealing, click the middle button at the top of Layout to make a radial tree. Here, there are a lot of options for displaying the tree that I just leave at default. One change I do make is to click on a branch with a support value over 0.700 to select it, and click Cartoon at the top to collapse the branches and keep sample labels. This is an example of a finished tree:

|  |
| --- |
|  |

For final editing I export the picture so I can add colors and shapes to the tree in PowerPoint. To export, select export graphic under File. Here, select Windows enchanced metafile (.emf) click options and check Draw text as shapes. Click Ok, name the file, and finish up by clicking Ok. Now that the tree is exported, you can finish up any editing in PowerPoint.