**NextGen Sequencing of the Grylloblattidae: Hypothesis Testing using Phylogenetic Analysis.**

Background: Schoville & Roderick (2010) did a preliminary phylogeny of the Grylloblattidae to try and determine how these creatures have moved around throughout history; their dataset had approx. 5400 bases of DNA data obtained with PCR. Results were somewhat ambiguous. Later, NextGen Sequencing was used to obtain a much larger dataset of over 208,000 bases of DNA from 323 different genes. Here, we explore how to build a tree using this large genomic dataset, and how to interpret the results of this tree. First, take a few moments to read the relevant sections (page 369) of Schoville & Graening (2013) and discuss some of the biogeographic hypotheses that have been, or could be, tested.

A. Tree building instructions:

1.  Download our genomic data (“gryllo\_genomic\_dataset.phy”) and open it in BBEdit or TextEdit and take a look at it.  It is in the “phylip” file format.  In this file format, there’s a header row with the number of taxa in the data matrix, followed by the number of characters in the matrix, separated by at least one space.  Subsequent rows have the taxon names followed by all the characters, which have already been aligned. This is a very common type of dataset in phylogenetic analysis.

2.  We will run a Maximum Likelihood search to obtain the most likely tree, given our data. The relatively-quick RAxML is a very popular software for this. Download the zip file of RAxML from (<https://github.com/stamatak/standard-RAxML>) using the green “download” button near the top, and double click on the zip file to unzip it. Open the Terminal (already installed on your computer, just search for it), type “cd” followed by a space, then drag the new RAxML folder into the terminal, and type enter.  This will have us inside the RAxML folder. Then type

make -f Makefile.gcc

to get RAxML up and ready to run (you may need to follow an additional prompt or two to download some additional software if this is your first time doing something like this).

3.  Make sure our phylip file is inside your RAxML folder.  Now, we will run our analysis in Terminal. Be sure you’re still working within the RAxML folder as before.  We will execute a run of 30 bootstraps (see <https://www.megasoftware.net/mega4/WebHelp/part_iv___evolutionary_analysis/constructing_phylogenetic_trees/statistical_tests_of_a_tree_obtained/bootstrap_tests/hc_bootstrap_test_phylogeny.htm> for an explanation of bootstrapping) using the “raxmlHPC” script and a “GTRGAMMA” model of molecular evolution (GTR specifies that each of the six possible transition and transversion rates are estimated separately; gamma specifies that each character can evolve at its own rate, independent of other characters); we will use the fastest version of a likelihood + bootstrap search (-f a); here you can see we also have random starting “seeds” for both our tree (-p) and 30 rapid-bootstrap trees (-x with any random number, then -# with 30); these specify the topology/branching pattern of a starting tree and any random number will work here.  In Terminal, type

./raxmlHPC -f a -m GTRGAMMA -p 12345 -x 12345 -# 30 -s gryllo\_genomic\_dataset.phy -n youroutputfilename.tree

where “youroutputfilename” is whatever you want to call your resultant tree, and let the analysis run.  A number of new files will be generated. You may see some warnings/errors too, but these are just alerting us to the fact that our data matrix has a few missing bases; this is nothing to worry about. At first, it may not seem like much is happening, but eventually updates will appear in the Terminal. Your analysis will be difficult for your computer to do and will probably take ~8 hours; when you’re not transporting it, please keep your computer open (you can dim the screen if you like) and plugged in as your analysis runs; make sure it is set not to go to sleep or turn off (temporarily change settings in System Preferences -> Energy Saver). 100 bootstraps would be more typical and better but it would take us too much time so we’ll live with 50. You’ll know your analysis is complete when you see something saying “Overall execution time for full ML analysis” followed by a number, and you’re returned to a new command line.

B. Tree visualization and annotation:

4.  Once your analysis is complete, we can visualize your best bootstrapped output tree using FigTree (download the latest version from the GitHub link found at <http://tree.bio.ed.ac.uk/software/figtree/>) and opening your newly generated file named “RAxML\_bipartitions.youroutputfilename.tree”.  FigTree will prompt you to name the node values, and you can name them “bootstraps” or something similar.  Play around with some of the FigTree options such as Reroot, Rotate, and various Layout options, and especially Node Labels, to make sure your tree is correctly rooted (I’d suggest using BK\_001 as your outgroup) and everything is legible and easy to read.  Note that the branch lengths of your tree correspond to the number of changes per site per period of time: The longer the branch, the more mutation has occurred leading to that taxon. Once your tree looks nice to you, print it out so that you can hand-write some geographical locations on it, or, you can save it as a PDF that you can polish in Illustrator or other program of your choice.

5. Using information in Table S3 in the Word file “SupplementalFile\_revision2”, add some geographical locality information next to the taxon codes. No need to write this for all 91 taxa, but maybe select a few from each major clade in the tree. Can you see any patterns emerge? For example, Schoville and Graening (2013) discuss that prior work using much smaller datasets suggests that Grylloblattidae may not have always simply moved from the north to the south in western North America, but what does this new dataset suggest? You may want to use Google Maps or Google Earth to see where some of the localities are, or even draw arrows suggesting movement of Grylloblattidae on a map.

6. Once you think you know the basic gist of what we’re trying to do here, and have a hypothesis for how Grylloblattidae have moved throughout history, you can compare your work to a finished product: Open Schoville et al 2018 and look at especially Figs. 3 and 5 to see how this all wrapped up after a slightly more robust analysis. For our purposes, Maximum Likelihood is roughly equivalent to Bayesian analysis, so your results should be pretty similar to what’s shown in this paper. Were they?