**Exercise 5: Estimating selection using Datamonkey**

EMD 531, Spring 2020

The purpose of this exercise is to explore models for estimating selection using Datamonkey – an open source collection of bioinformatic tools: <http://www.datamonkey.org/>

***\*\*Important***: Before class, make sure that you have downloaded **ZIKV\_test-data\_aligned-trimmed.fasta** from the Test Data folder. We will practice in class using our test dataset, but you will need to complete the following with your own dataset to complete the exercise. We will use one branch and two codon site models for inferring the strength of natural selection in your data.

For your aligned + trimmed .fasta file, **ensure that the stop codons are removed!** Datamonkey will not run with stop codons.

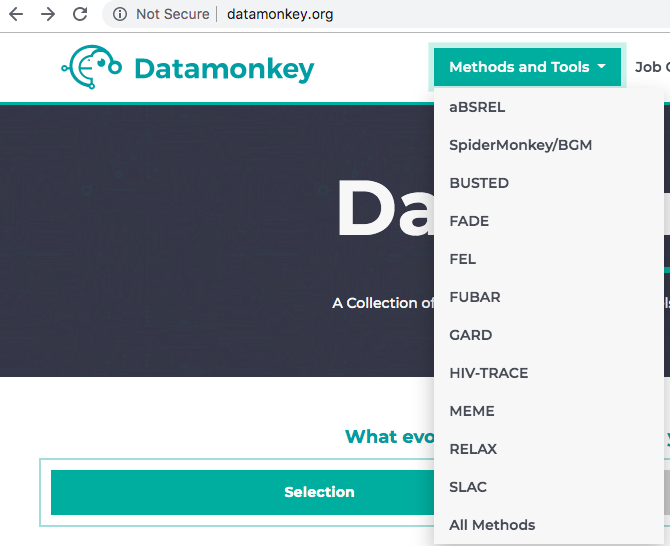
Tests for selection:

|  |  |
| --- | --- |
| ω | Direction of selection: ~ 1 neutral evolution; < 1 negative selection; > 1 positive selection |
| β | Rates of non-synonymous substitutions (dN) |
| α | Rates of synonymous substitutions (dS) |
| LRT | Likelihood Ratio Test; tests if *dN* is significantly greater than *dS* |

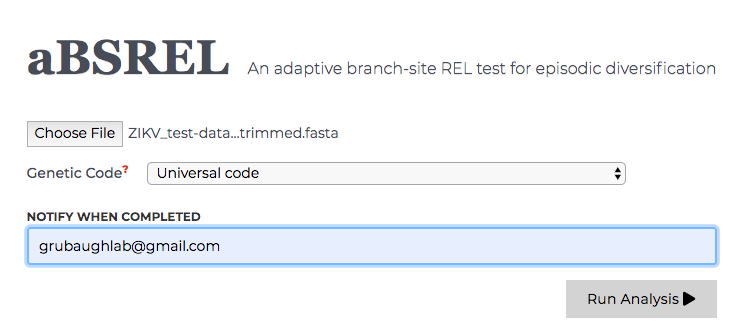
**Overview:**

**Step 1:** Branch model assuming episodic selection: aBSREL (adaptive Branch-Site Random Effects Likelihood)

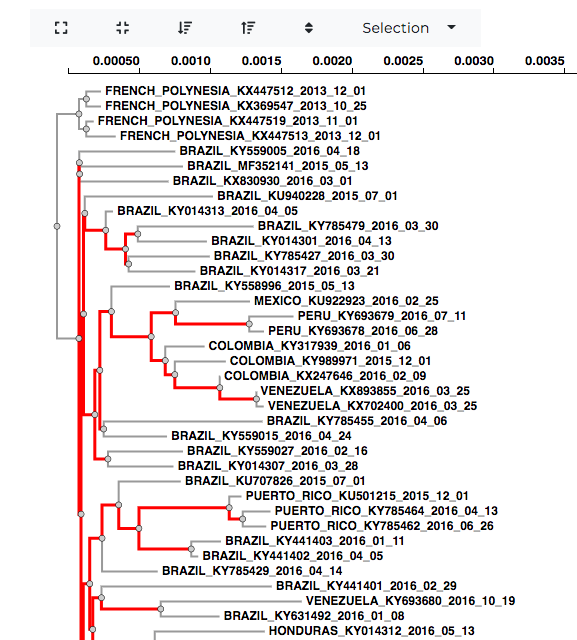
* aBSREL is a "branch-site" model to test if positive selection has occurred on a proportion of branches (diversifying selection). It will provide an ω score for each branch and assess significance using the Likelihood Ratio Test at a threshold of p ≤ 0.05 after correcting for multiple comparisons It will then report evidence of episodic diversifying selection in your phylogeny. aBSREL can be run in two modes:
  + 1) Hypothesis testing in which the user selects a set of branches a priori to test for positive selection.
  + 2) Exploratory analysis in which all branches in the phylogeny are tested for positive selection. Due to more comparisons, the exploratory approach has much lower power.
  + See here for more details: <http://hyphy.org/methods/selection-methods/#absrel>
  + Paper: <https://academic.oup.com/mbe/article/32/5/1342/1130440>
* To run aBSREL, follow this Datamonkey (<http://www.datamonkey.org/>) path:
  + **Selection > Branches > Episodic**
  + Or select **aBSREL** from the Methods and Tools dropdown menu



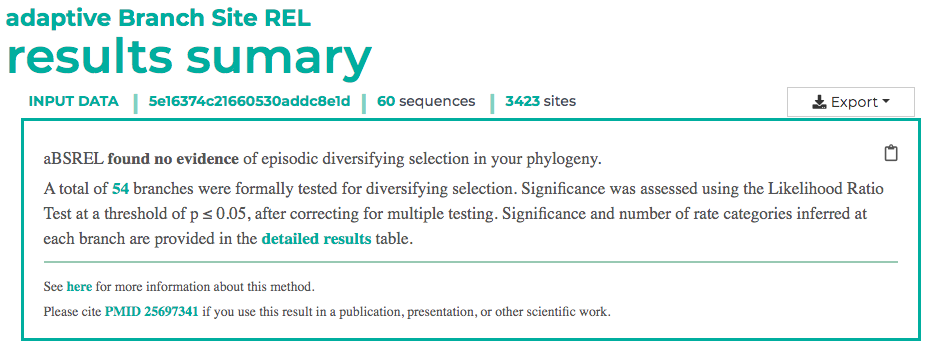
* Choose the **ZIKV\_test-data\_aligned-trimmed.fasta** file, select **Universal Code** for Genetic Code, and **Run Analysis**.

****

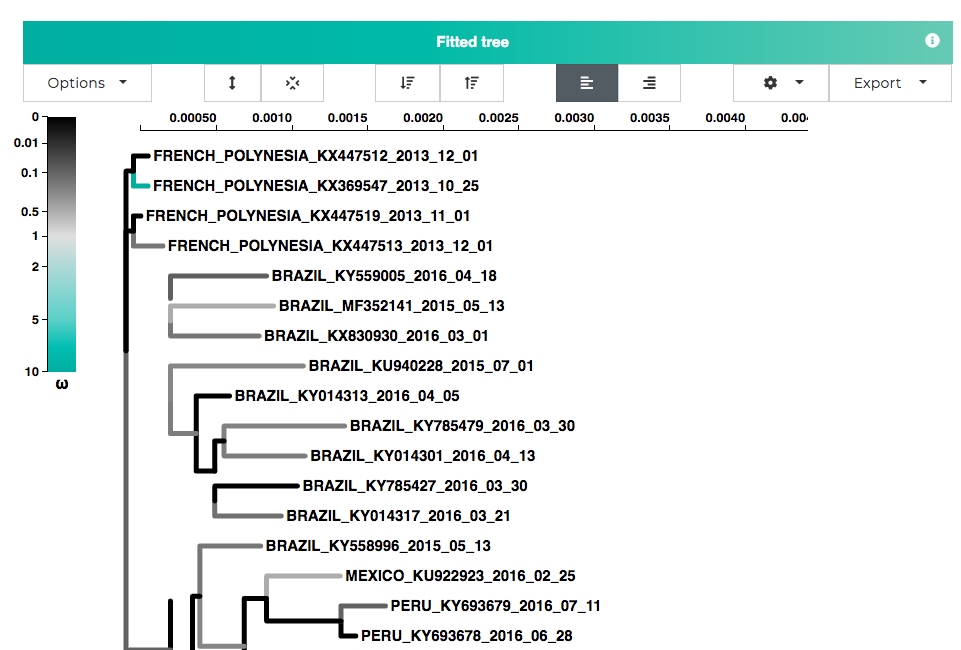
* The next screen will show an interactive phylogenetic tree of your sequences to select the test branches. Before running, perform the following steps:
  + Properly root the tree by selecting the node – in this case the node that contains all of the French Polynesian sequences – and selecting **reroot on this node**.
  + Select the test branches by selecting the node that contains all the test sequences – in this case the node that contains all of the sequences from the Americas – and selecting **All internal branches**. This will allow us to only test selection on ancestral sequences that have had adequate time for selection.



* Run the program by selecting the **Save Branch Selection** at the bottom of the screen to run. The job will take several minutes to complete. Datamonkey is a web-based server, run time can depend on the number of people using it at one time. If you entered your email on the previous page, you will get an alert when it if finished.
* When the run is finished, it will provide a results summary. In this case, no diversifying selection was found. Take a screenshot for question **E5-1** below.

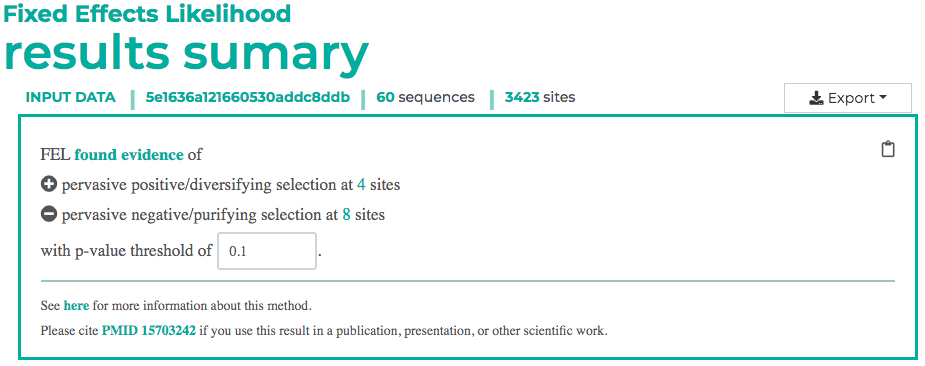
****

* More detailed results, including a tree containing ω scores for each branch and a detailed results table, are found below. Download a **PNG** of your Fitted Tree using the **Export** dropdown menu for question **E5-1**.

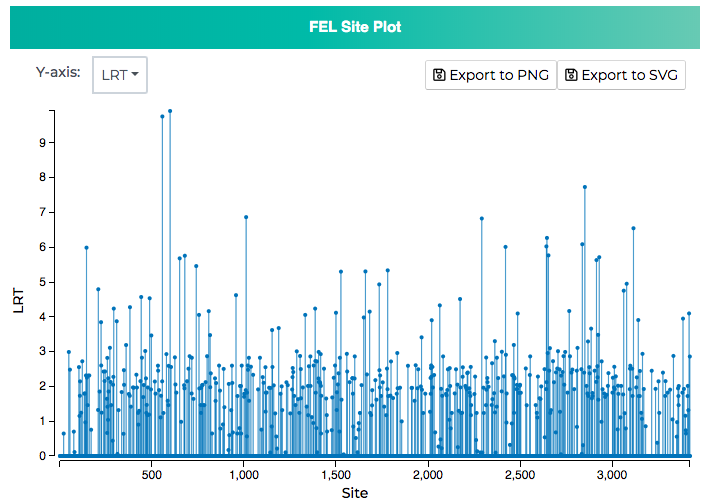
****

**Step 2:** Site model assuming pervasive selection: FEL (Fixed Effects Likelihood)

* FEL uses a maximum-likelihood approach to infer nonsynoymous (*dN*) and synonymous (*dS*) substitution rates on a per-site (codon position) basis for a given coding alignment and corresponding phylogeny. This method assumes that the selection pressure for each site is constant (pervasive) along the entire phylogeny. Hypothesis testing is then conducted on a site-specific basis, using the Likelihood Ratio Test (LRT), to ascertain if dN is significantly greater than dS. Like aBSREL, FEL can be ran in 2 modes (all or specific branches).
  + See here for more details: <http://hyphy.org/methods/selection-methods/#fel>
  + Paper: <https://academic.oup.com/mbe/article/22/5/1208/1066893>
* To run FEL, follow this Datamonkey (<http://www.datamonkey.org/>) path:
  + **Selection > Sites > Pervasive > Small**
  + Or select **FEL** from the Methods and Tools dropdown menu
* Choose the **ZIKV\_test-data\_aligned-trimmed.fasta** file*,* select **Universal Code** for Genetic Code and **Yes** for Synonymous rate variation, then **Run Analysis**
* Reroot your tree and select test branches (internal branches) as outlined in the aBSREL step. Run by selecting the **Save Branch Selection** at the bottom of the screen to run.
* When the run is finished, it will provide a results summary. Record the number of sites where FEL found pervasive positive/diversifying selection and pervasive negative/purifying selection at p-value thresholds of 0.1 and 0.05.

****

* More detailed results, including a FEL Table and FEL Site Plot are found below.
* Sort the FEL Table by highest **LRT** and record the 1) site, 2) alpha, 3) beta, 4) LRT, and 5) p-values for the top 5 positively and negatively selected sites (as ranked by LRT scores, so 10 sites in total). If you do not have at least 5 sites for each, then add more from the other category to reach 10 sites in total.
* Select **LRT** for the Y-axis of the FEL Site Plot and Export a PNG for question **E5-2**.



**Step 3:** Site model assuming episodic selection: MEME (Mixed Effects Model of Evolution)

* MEME employs a mixed-effects maximum likelihood approach to test the hypothesis that individual sites have been subject to episodic positive or diversifying selection. In other words, MEME aims to detect sites evolving under positive selection under a proportion of branches (so just from part of the tree). Not accounting for episodic selection vastly underestimates the number of sites experiencing positive selection.
  + For more information: <http://hyphy.org/methods/selection-methods/#meme>
  + Paper: <https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002764>
* To run MEME, follow this Datamonkey (<http://www.datamonkey.org/>) path:
  + **Selection > Sites > Episodic**
  + Or select **MEME** from the Methods and Tools dropdown menu
* Choose the **ZIKV\_test-data\_aligned-trimmed.fasta** file, select **Universal Code** for Genetic Code, and **Run Analysis**. There is no test branch selection this time.
* When the run is finished, it will provide a results summary. Record the number of sites where FEL found pervasive positive/diversifying selection at p-value thresholds of 0.1 and 0.05.
* More detailed results, including a MEME Table and MEME Site Plot, are found below.
* Sort the MEME Table by highest **LRT** and record the 1) site, 2) alpha, 3) beta, 4) LRT, and 5) p-values for the top 5 positively selected sites (as ranked by LRT scores).
* Select **LRT** for the Y-axis of the MEM Site Plot and Export a PNG for question **E5-2**.

**Questions:**

*Combine answers with all exercises,* ***due before class 16 on March 5th****. Be prepared to discuss your answers during class 16.*

**Rerun the steps about using your data, not the ZIKV test. Make sure that you reroot your tree and select internal branches as it makes sense for your data. Then answer the following questions.**

**E5-1**: Paste a screenshot of your **aBSREL** results summary and a PNG of your fitted tree. (3 points)

**E5-2**: List the number of sites where **FEL** found pervasive positive/diversifying selection and pervasive negative/purifying selection at p-value thresholds of 0.1 and 0.05. Create a table that includes the 1) site, 2) alpha, 3) beta, 4) LRT, and 5) p-values for the top 5 positively and negatively selected sites (as ranked by LRT scores, so 10 sites in total). If you do not have at least 5 sites for each, then add more from the other category to reach 10 sites in total. (3 points)

**E5-3**: List the number of sites where **MEME** found pervasive positive/diversifying selection at p-value thresholds of 0.1 and 0.05. Create a table that includes the 1) site, 2) alpha, 3) beta, 4) LRT, and 5) p-values for the top 5 positively selected sites (as ranked by LRT scores). (3 points)

**E5-4**: Looking at your combined results, what results are consistent across the tests? What results are different? For the two site models, where does most positive selection occur within the genome or gene? (Max 200 words; 3 points)

**E5-5**: Select one positively selected site you identified in your top LRT list (preferably a site that is identified by both FEL and MEME). Open your sequence alignment in UGene and find the site that you selected. Remember, the sites are numbered by codon position, not nucleotide (e.g., codon site 107 refers to nucleotide positions 319, 320, and 321). Find the sequences that have a substitution compared to the root sequences at the site. Next, find these sequences in your Nextstrain tree, highlight the position or clade, and paste a screenshot of your tree and map. Based on where they sit in your tree, can you make a hypothesis about the selection pressure and potential phenotype of the substitution? How would you experimentally test this hypothesis? (Max 300 words; 8 points)