Group B Final Project Tool

Andrew Meng, Kathi Munoz-Hofmann, Joseph Zou

Calculating and Visualizing Fst:

This Python program serves as a tool for calculating and visualizing FST values at each segregating site of the given population data. FST serves as a measure to compare heterozygosity between sub- and total populations at a given segregating site. The equation for FST is

$$Fst = 1 - \frac{Hs}{Ht} .$$

Where *Hs* indicates heterozygosity of a segregating site in a given subpopulation and *Ht* represents heterozygosity of that same site for the total population.

This tool provides a graphical representation of these Fst values in order of the sites for each provided subpopulation. Additionally, upper and lower bounds for statistical outliers of these values are calculated and represented on the same graph. Thus, statistical outliers can be easily picked out for a subpopulation. We determine outliers to be those that exceed 1.5 times the interquartile range of the calculated Fst values. Upper and lower bounds for outliers are then calculated as follows:

1. Determine the 1st and 3rd quartiles (call them q1, q3). The tool uses the python library numpy's 'percentile' function to do this.
2. Determine the interquartile range (*iqr*) by taking the difference of the 1st and 3rd quartiles:

$$iqr = q3 - q1 .$$

3. Determine the upper (*ub*) and lower (*lb*) bounds for outliers as follows:

$$ub = q3 + 1.5 * iqr ,$$
$$lb = q1 - 1.5 * iqr .$$

The tool then makes use of Python's graphical library matplotlib to create graphs for each subpopulation presenting both the F_{st} values for each subpopulation and their corresponding outlier indicators.

How to install the program:

The tool is located on the course git repository at:

ComputationalGenetics/Group B/Star_Wars/

And is called 'FSTtool.py'.

In order to run this program, the user needs to install matplotlib and numpy beforehand using commands:

```
pip install matplotlib
pip install numpy
```

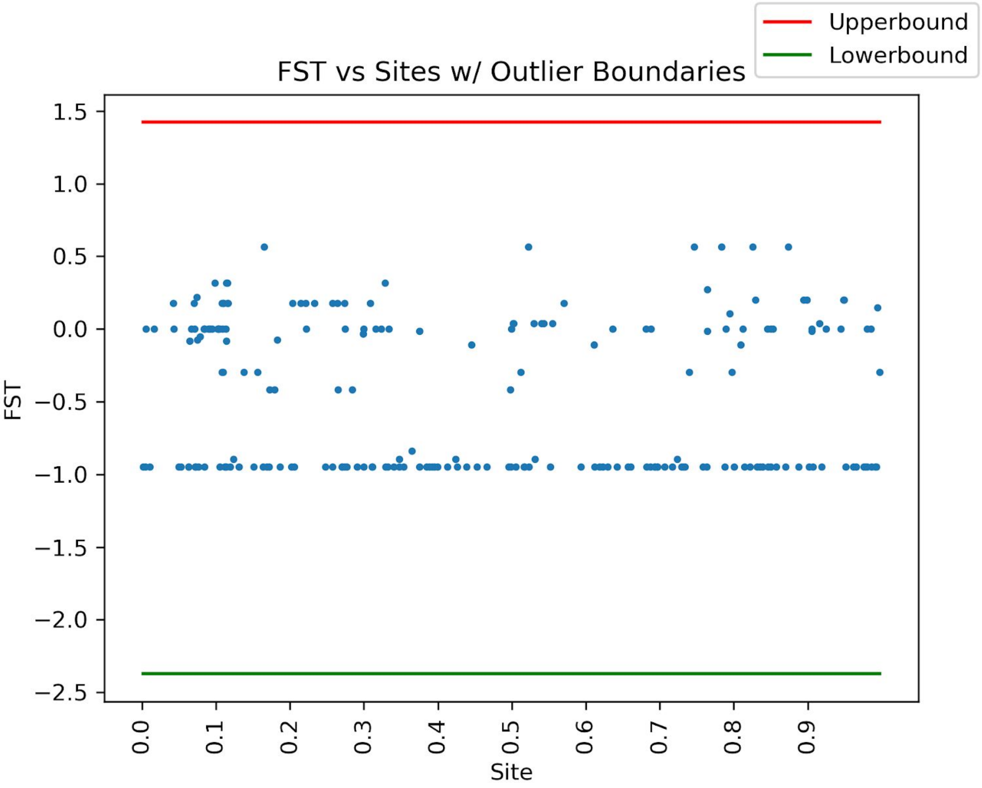
Python 3 is required for this program to run correctly. Apart from the external library required, there is no installation required for the actual program.

What the input needs to look like:

The input to this program should be a set of files output from SLiM. The output is of the form generated by SLiM's outputMS method. The first argument to the program is the name of the folder to which output will be written; each subsequent argument to the main method is a file name, and each file should contain a description of a sample of a subpopulation containing only segregated sites. The first file should be a description of the total population, and each additional file should describe a subpopulation. The sample included in the first file should contain the same individuals as those in subpopulations.

What the output will look like:

The output of the program will be some graphs saved as PNG files in the folder specified in the input. The PNG files will share the names of the input subpopulation files. The image would look like this:



One graph per subpopulation will be produced. The y-value of a dot on the graph represents the FST calculation at the site described by the x-value of the dot.

<u>How to run the tool:</u>

Use the command line:

```
python3 FSTtool.py <folder name> <total population> <subpopulation 1> <subpopulation 2> <subpopulation 3> …
```

Example: python3 FSTtool.py temp Humans-total.txt Human-p6.txt SubHum-p60.txt

If the python3 command is not available, try python instead.