Using the Python Comparison Script Tool

Description: This script consolidates and permits the comparison of forensic reports generated with Illumina’s Universal Analysis Software (\*.xls), ThermoFisher Scientific’s HID SNP Genotyper (\*.csv), and developed CLC workflow (\*.tsv). It is capable of analyzing any number of the reports side by side and it exports data on the coverage, genotype, flag information, and major allele frequency for 301 forensically relevant ancestry and identity SNPs.

Dependencies: This script was developed using Python v3.5.1. Both openpyxl and xlrd need to be loaded into python before executing, as these modules are needed to access excel files of different formats (these modules can be installed using a tool such as pip; they are not included in the supplemental files for this workflow). The included reference file, *loci.py*, which contains the SNP location information needs to be placed in the same directory as this script (comparisonScript.py).

Preparation: Place all report files that are to be processed into a single folder. Ensure that the naming syntax for files from a single sample is consistent across report files (e.g. Sample1.csv; Sample1.xls, Sample1.tsv).

Execution: An example of how to run the script from the terminal is as follows:

C:\python comparisonScript.py -c -g -f -m -n -i D:\Project\_A\_reports -o D:\Project\_A\_Reports\project\_A

\*This command will take all files found in the Ancestry folder that have an extension of \*.csv, \*.tsv, \*.xls, or \*.xlsx and create an output file named project\_A\_SNPs.csv in the Project\_A folder. The report will include all four attributes: genotype, coverage, flags, and major allele frequency.

The following parameters may be used to tweak the output to the users preference:

* -c - coverage information will be included in the report
* -f – flag information will be included in the report
* -g – genotype information will be included in the report
* -h – help (displays an example command line entry)
* -i – tells the script where the parent folder is that contains all files to be processed
* -m – major allele frequency information will be included in the report
* -n – will include a file that contains all SNP information (postfixed as \_SNP.csv)
* -o – Tells the script the location where to store the output file(s) and will use the final folder identifier as a prefix for the output file