Legal Aid Statistical Tool

The Legal Aid Society in New York City

This tool can be used to achieve likelihood ratio for suspect(s) and/or victim(s) in a DNA mixture using probabilistic genotyping

Legal Aid Statistical Tool Copyright (C) 2015 Munieshwar Ramdass

This program is free software: you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation, either version 3 of the License, or (at your option) any later version.

This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.

You should have received a copy of the GNU General Public License along with this program. If not, see <http://www.gnu.org/licenses/>.

Legal Aid Statistic Tool (LAST)

# Introduction

The Legal Aid Statistical Tool (LAST) was made to replicate some of the functionality of the Office of Chief Medical Examiner’s Forensic Statistical Tool (FST). Both LAST and FST are probabilistic genotyping tools which can be used to determine the likelihood ratio of a suspect and/or victim in a sample of DNA mixture after the mixture has gone through the electropherograms. No attempt was made on LAST to improve the statistical analysis as its purpose was served as a replicate of FST. LAST’s logic was implemented by Munieshwar “Kevin” Ramdass in C++ using Visual Studio. LAST’s graphical user interface was implemented by Nicholas J. Corpuz and Khagay Nagdimov in R using R Studio. LAST was a summer of 2015 project for this team of interns at the Legal Aid Society.

# Functionalities

As of the August 5, 2015, LAST can perform the following functions provided they are accurately imputed into a CSV file. It is possible for LAST to perform more combinations with larger number of contributors providing there is an acceptable drop-out rates table for that number of contributors.

LAST can potentially perform analysis on mixtures that include more than four persons providing the computer it is run on is capable of doing it. Note that analysis of mixtures with more than three persons are very time and space consuming for computers to perform.

# System Requirements

To be able to Run LAST, the follow minimum hardware and software are required.

Hardware Requirements

CPU: 2.27 GHz or more

RAM: 4GB or more

Disk Space: 8GB or more

Software Requirements

Microsoft Windows 7 or more recent version of Microsoft Windows

Microsoft Office Excel 2013 or more recent version of Microsoft Office Excel

# Manual Interface

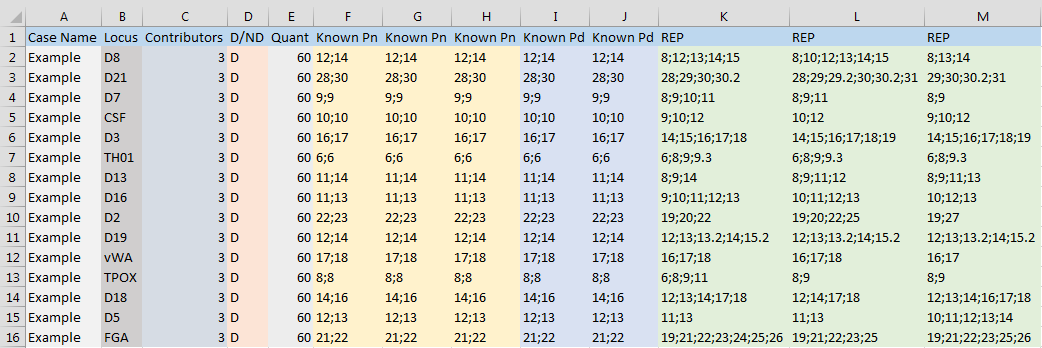


Figure 4.1: An unrealistic case.csv

Ideally, LAST will be run with 15 loci from a specific case. In a “.csv” file named “case.csv”, which can be created using Microsoft Office Excel 2013 or later, create the following table starting at cell A1. The first row must contain the header that is specific to the case. Headers may include as much “Known Alleles” columns consecutively as there are “Contributors” specified. There may be no “REP” columns to the far right or as much “REP” columns to the far right consecutively as needed.

The first column is the case name. Ideally there should be one case name for all of the 15 loci though this is not a requirement.

The second column, “Locus,” is the specific locus. Ideally there should be 15 different loci for a total of 16 rows in this CSV file to run a full case and achieve an accurate overall likelihood ratio from LAST. Locus names must be spelt as shown above.

The third column is the “Contributors” column which is the number of knowns plus unknowns. A sample with one known and two unknowns is a three person mixture. The number “3” will be inputted under “Contributors” and it should remain constant for all rows if a full case is being run.

The fourth column is “D/ND” where “D” or “ND” or “d” or “nd” or “yes” or “no” may be entered to specify whether a run is deducible or not. This should remain constant for all 15 rows in a full run.

The fifth column is “Quant” where a numerical value is inputted to specify the quant that the case is run at. This should remain constant for all 15 rows in a full run.

The following column(s) are the “Known Pn” columns where there must be two numerical values inputted separated by a semicolon. This column is for known contributors that belong in the numerator only. More “Known Pn” column may be inserted directly to its right so as long as it does not exceed the number of “Contributors.” There must be at least one “Known Pn” column.

After the “Known Pn” columns, are the “Known Pd” columns where there must be two numerical values inputted separated by a semicolon. This column is for known contributors that belong in the denominator only. More “Known Pd” column may be inserted directly to its right so as long as it does not exceed the number of “Contributors.” Is it possible to have no “Known Pd” columns.

The last few columns are “REP” columns for each replicate from the electropherograms which contain numerical values separated by semicolons. Enter “INC” if a replicate is deemed inconclusive or “NEG” if a replicate or run is condemned. Usually for an “INC” run, all replicates that are affected should be marked with “INC” to ignore that replicate during the run. Usually for a “NEG” run, all “REP” columns will be marked with “NEG” to achieve a likelihood ratio of 1 for the locus.

The following is an unrealistic run to show what LAST can handle as input:

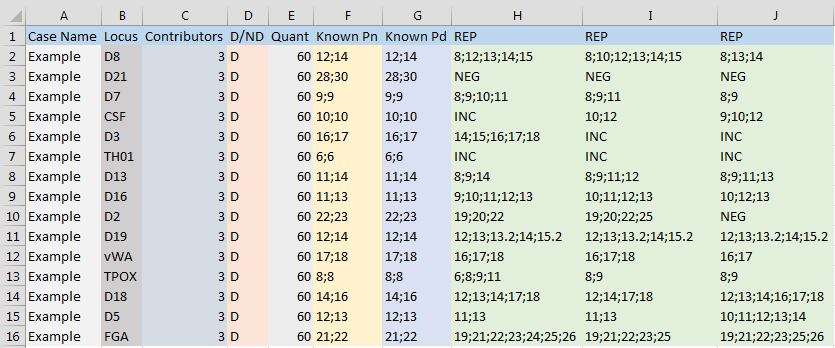


Figure 4.2: An unrealistic case.csv

The Evidence\_File\_Generator.exe will generate 15 CSV files that will be named “Evidence\_1.csv” to “Evidence\_15.csv” using the “case.csv” file described above. Files created are determined as number of rows in the “case.csv” file. Do not enter allele lengths that simply do not exist.

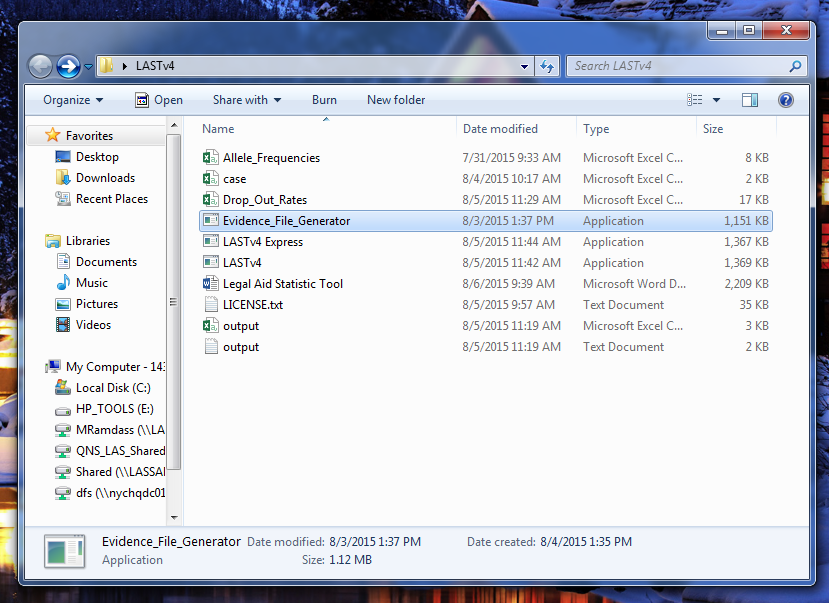


Figure 4.3: Directory that all files are in.

Double Clicking on “Evidence\_File\_Generator” as shown above will generate 15 files as shown below. Note that the “case” CSV file must be in the same directory at the “Evidence\_File\_Generator” in order to run.

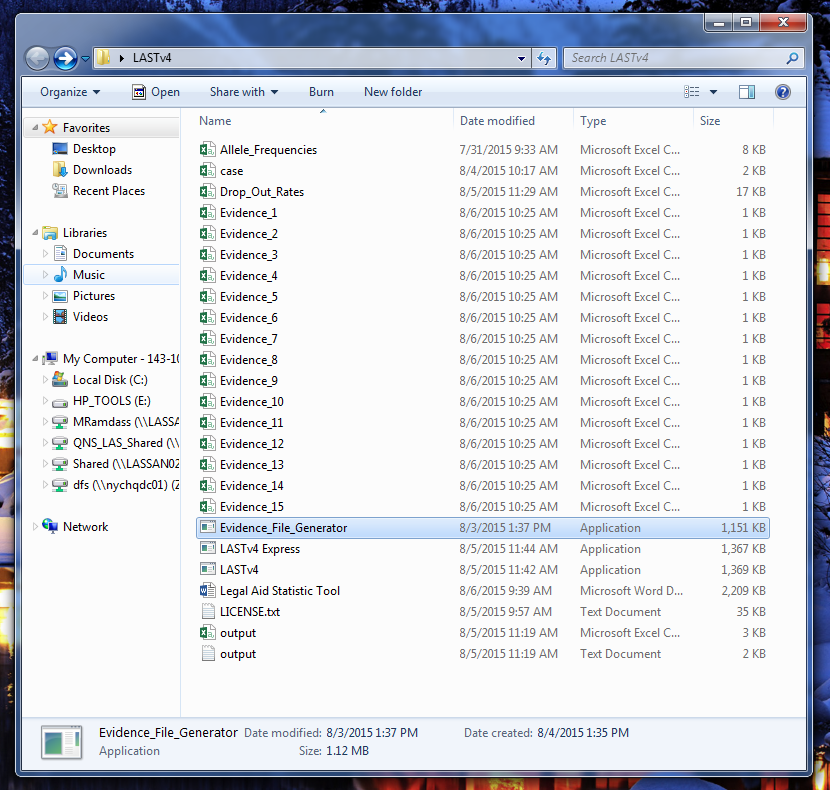


Figure 4.4: Directory all files are in now with evidence files.

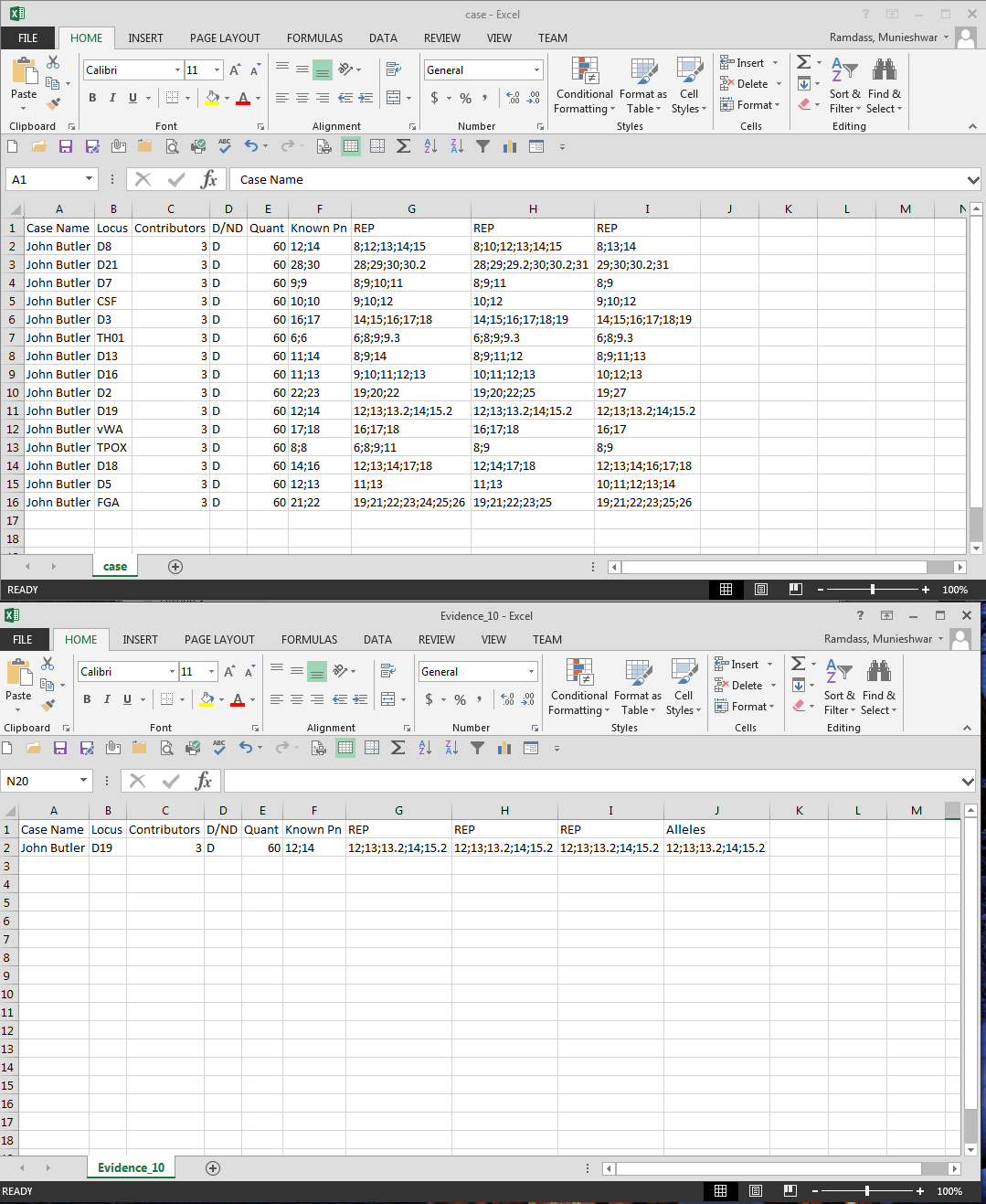


Figure 4.5: case.csv and Evidence\_10.csv

Note above how an “Alleles” column is generated in the “Evidence\_10” CSV file. This contains the unique alleles observed from the replicates.

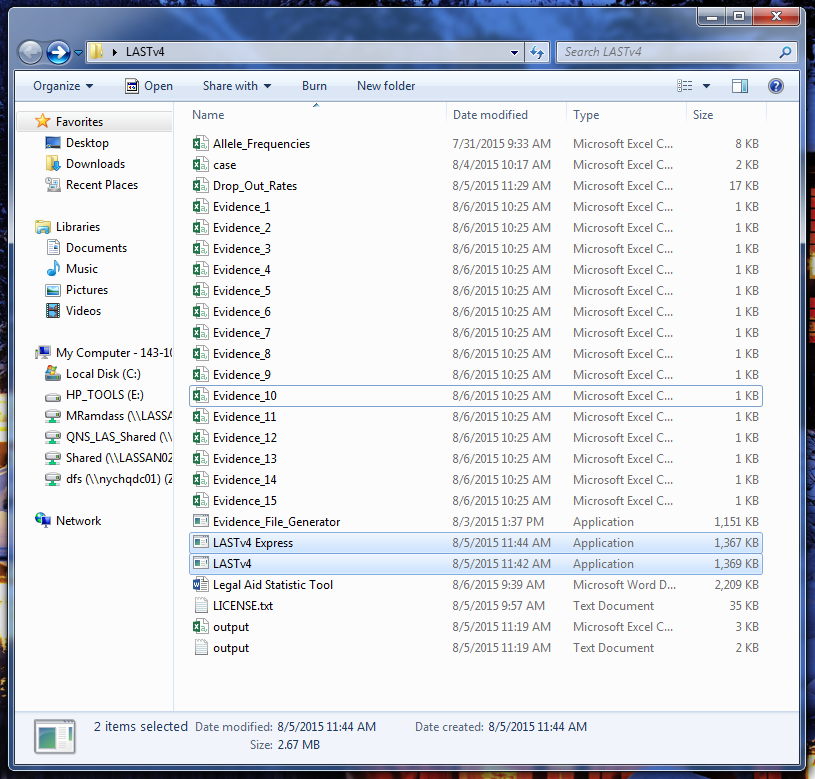


Figure 4.6: Directory that all files are in.

Next double click “LASTv4” as shown above. This will append data to each of the evidence files with a detailed progress shown in command prompt and create two outputs named “output.csv” and “output.txt” for formatting convenience as shown below. Aborting run can be done by exiting the command prompt. Note that if a rerun were to be done, start again by double clicking “Evidence\_File\_Generator” and continue the process described thus far. Note that in all runs the two files named “Allele\_Frequencies” and “Drop\_Out\_Rates” must be present in the same directory as the rest of the files. Alternatively, you may click LASTv4 Express with will run faster but will not append the matrix used to show the values being multiplied and added to achieve the likelihood ratio.

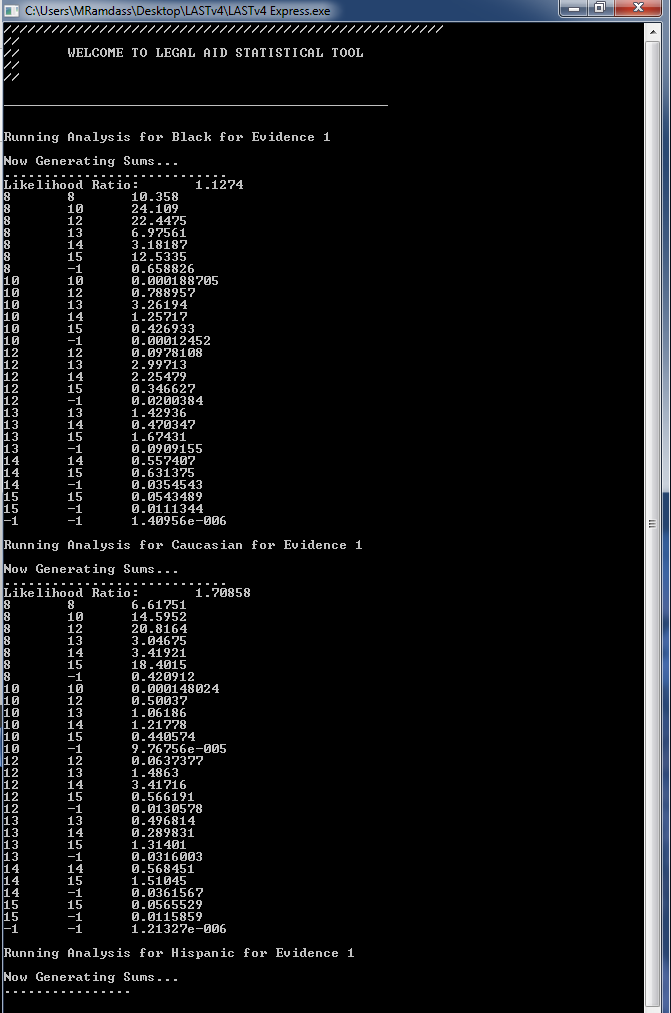


Figure 4.7: Command prompt showing running status.

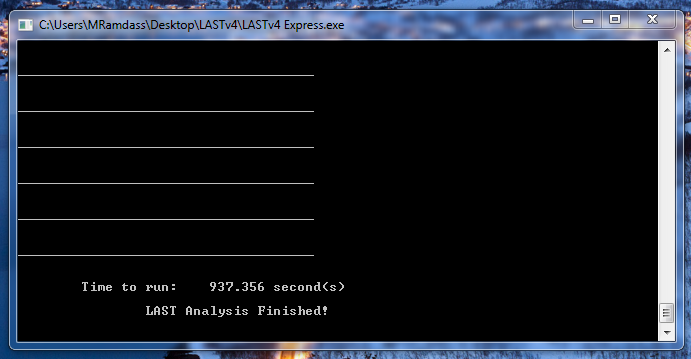


Figure 4.8: Command prompt indicating LAST run has finished and the time it took to run.

Refer to Viewing Data to learn about analysis.

# User Interface

There may be several interfaces developed for this software. Refer to Legal Aid Statistical Tool User Interface documentation for the interface you wish to use.

# Viewing Data

When LAST indicates that it is done, hit enter in the command prompt or exit it. And view the outputs and/or evidence files.

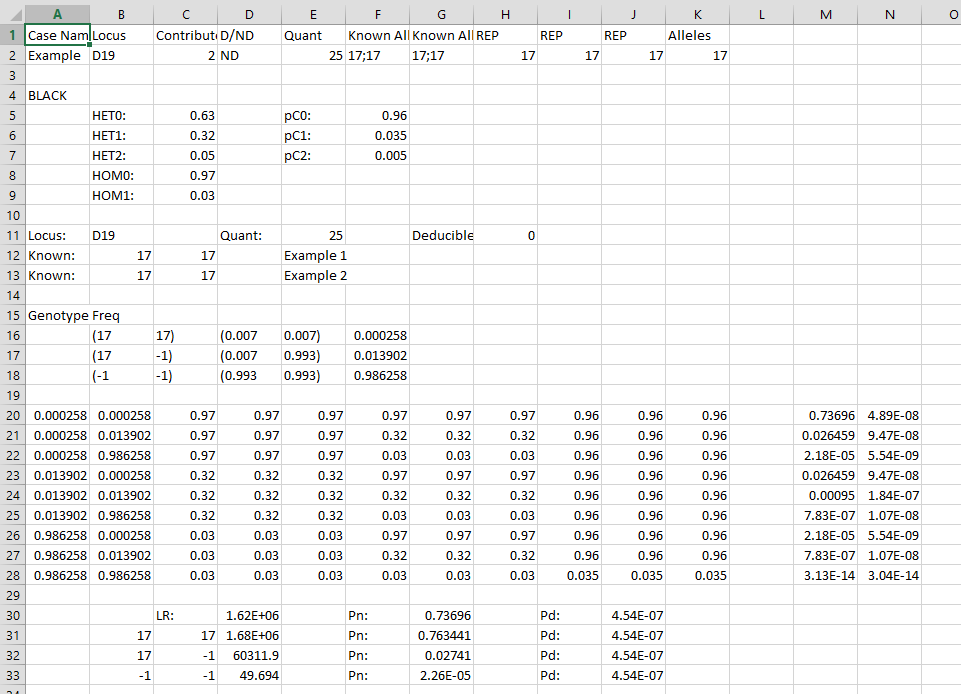


Figure 6.1: Evidence\_1.csv and manual calculation shown.

Above is from “Evidence\_1.csv” which shows the appended data for manual calculations. Combinations or LRs shown are with respect to the first “Known Alleles” column. However, the most important LR for the run is the first LR shown which will be used to calculate the overall LR. In the block (or matrix – rows 20 to 28) of calculations above, the first two columns are genotype frequencies for every combination in the population for this locus and case. The following columns (C to H) are in groups of two (two because “Contributors” is “2”), first being the drop-out rates for the first known and second being the drop-out rates for the second known. From column I to K are drop-in rates for each replicate respectively. Other values used are specified above, for example, under “Genotype Freq” labels the allele lengths 17 and 17, then their respective locus frequencies of 0.007 and 0.007, then lastly their genotype frequency of 0.000258. This is done for each race, Black, Caucasian, Hispanic, and Asian.

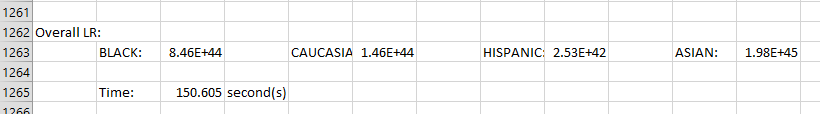


Figure 6.2: output.csv and overall likelihood ratios for each sub population shown.

Both “output” files show the same information in a different format. The overall likelihood ratio is shown above for the entire run using the 15 evidence files. A brief report of each run is also given. Said report is similar to the data in the evidence file; it is without the calculation matrix for each evidence file.

At the end of each run of LAST, move all the evidence files and output files elsewhere as they will be rewritten the next time “Evidence\_File\_Generator” or “LASTv4” or “LASTv4 Express” is run.

# Use and Future of LAST

LAST will be used for case studies at the Legal Aid Society and to compare to FST runs as well as other probabilistic genotyping software. The user interface and logic made be changed in the future if bugs or incorrect methodology has been observed through running LAST. LAST may update to reflect changes in FST also. There may be several user interface for LAST. The interfaces described in this document are from the summer of 2015 intern team of Munieshwar Ramdass, Nicholas J. Corpuz and Khagay Nagdimov.