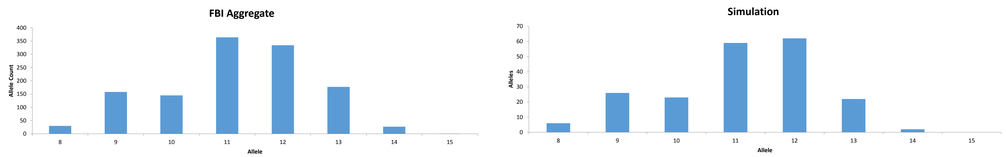
# Simulating Families from STR Data Derived from Files Exported from CODIS

Daniel Myers, New York State Police Forensic Investigation Center

To measure the sensitivity and specificity of a STR offender database for familial searching calculations, obtaining an appropriate sample size of true families may not be a viable option. A possible solution is offered to create biological, simulated family members from offender samples (i.e.; seeds) to be used in familial searching validation and assessment of likelihood ratio cutoff values specific to the database being searched. The simulation utilizes Excel VBA macros to import CODIS XML (Extensible Markup Language) files in Interpol format into memory as Microsoft Document Object Models (DOMs). The simulation creates DOMs for two biological parents, two biological full-siblings and two biological half-siblings (one from each parent) with simulated alleles chosen according to an aggregate of the amended FBI Caucasian, African-American and Southwestern Hispanic frequencies [1]. The DOMs are then saved as XML files in Interpol format for import into familial searching software.

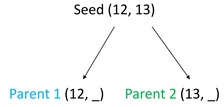
## Comparing distributions

To ensure the simulator is following the aggregate FBI distribution to select alleles, two distributions are automatically plotted with each simulation (one from a table containing the aggregated FBI allele sets and one from a table of the simulated alleles). According to previous population studies, samples from 100-300 individuals are adequate to approximate a population [2]. Below is a comparison of the distribution of alleles of the FBI aggregate vs a simulation of 800 alleles (equivalent of 400 individuals) at the D16S539 locus sampled from the FBI aggregate. All other locations also have distributions comparable to the aggregate (not shown).

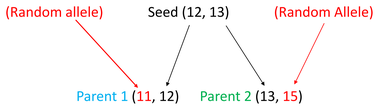



## Reverse Engineering of Parents

In the below example, the offender seed is a (12, 13) at D16S539 (shown in black). The simulator reads the 12, assigns it to the first parent (shown in blue) and then the 13 and assigns it to the second parent (shown in green).

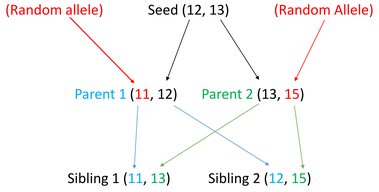



To simulate the missing alleles (shown in red), they are chosen from the aggregate FBI distribution using the random number method in VBA.



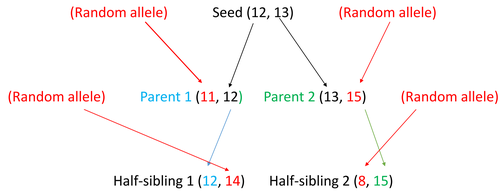

## Creating Full-siblings

To create full-siblings to the seed, the random number method again is used again but this time to select which allele from each parent will be used to create a new individual. This crosses the parents in Mendelian fashion (one possible outcome shown).




## Creating Half-siblings

To create half-siblings, the random number method selects one allele from each parent as before but instead of crossing for the second allele, the random number method uses the aggregated FBI amended distribution to simulate a mating with a random person from the population. In effect, this creates half-siblings to the seed (one possible outcome shown).

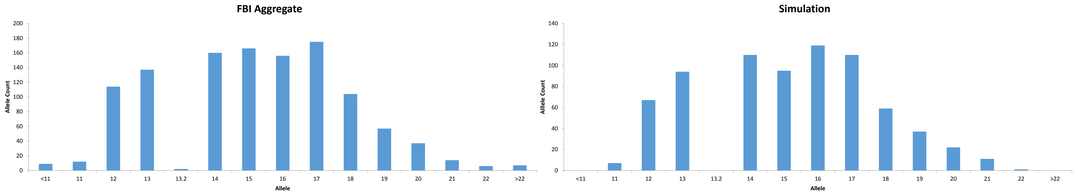



## Creating Large Numbers of Families

CODIS XML Interpol format files can contain any number of offenders in a single file. The usefulness in the simulator becomes apparent when many profiles are present. One file can be loaded into memory and an XML file can be created for each type of relative to the seed. The macro will iterate as many times as there are profiles in the XML file and iterate for every locus present creating full genotypes of simulated biological relatives to the seed.

## New York State Police Familial Searching Validation

For the New York State Police familial searching validation study, the macros were tested using 20 thousand Identifiler genotypes prior to conducting a sensitivity and specificity study of the New York State Convicted Offender Index. Below is a comparison of the distributions from the validation study of the FBI aggregate vs the simulation of 20 thousand alleles at the D18S51 locus. All other locations also have distributions comparable to the aggregate (not shown). For the sensitivity and specificity study 3,350 biological relatives of 1,150 offender seeds were simulated.




## Conclusion

Simulating large sets of biological relatives to offender seeds has two advantages. First, it removes the need to find true families willing to donate samples. It's more economically viable and time saving than attempting to find willing participants and performing lengthy laboratory work. Second, it allows for direct measurement of the database in question by using samples from it rather than adding more samples to it or simulating a database. This will become important in the future as the size of the database increases.

In addition, using the document object model greatly enhances the utility of the simulator. Instead of working with spreadsheets, XML files can be directly read, created and saved with more efficiency. This also allows for use in different programming languages. There would be more utility in a program written in language such as Python, VB, R or Java. Instead of a requirement of setting up Excel VBA Macros on each machine an executable file could be added to the machine to create a stand-alone program.

The simulators were found to adequately follow the aggregate FBI distribution and create biological relatives to offender seeds, allowing an estimate of the sensitivity and specificity of the New York State Offender Index for familial searching. Additional work is intended to make the simulator easier to use and estimate sensitivity and specificity of the growing database, including the expanded CODIS 20 core loci.

References
1. Moretti, T., Budowle, B.; Buckleton, J.; "Population data on the thirteen CODIS core short tandem repeat loci in African Americans, US Caucasians, Hispanics, Bahamians, Jamaicans, and Trinidadians; Erratum," *Journal of Forensic Sciences,* no. 4, pp. 1114-1116, 3 June 2015.
2. R. Chakraborty, "Sample Size Requirements for Addressing the Population Genetic Issues of Forensic User of DNA Typing," *Human Biology,* vol. 64, no. 2, pp. 141-159.