GGETIt: A Simulation Tool for the Generation and Evaluation of Genotypes Technical Manual

Developed and Written with Contributions by

Sarah Norsworthy
Catherine Grgicak

Boston University School of Medicine, Program in Biomedical Forensic Sciences, 72 East Concord Street, Boston, MA 02118

Table of Contents

1. Introduction

GGETIt is a simulation tool designed to generate single-source and mixture profiles with STR loci (and amelogenin) consistent with the AmpFℓSTR® Identifiler® Plus (Applied Biosystems®, Foster City, CA) and GlobalFiler™ (Applied Biosystems®, Foster City, CA) amplification kits for up to six contributors based on allele frequencies in the population [1-2]. Users input desired allele frequencies, the true number of contributors (NOC) to the profiles, the number of profiles they wish to generate, and the probability of dropout (Pr(D)) of each contributor. This simulation provides the end-user with the profiles, or known genotypes, for each contributor, color-coded to delineate the alleles per contributor that did (red) and did not (black) dropout. It also provides the 'observed allele output', which is representative of the alleles that would be detected in the electropherogram, and calculates the minimum NOC using the maximum allele count (MAC) method of the 'observed allele output'. The MAC method involves counting the number of alleles at each locus, dividing the maximum observed by two, and rounding up [3]. Three results tables are also provided: the true versus observed number of contributors for each profile (Table 1), the number of observed alleles per locus for each profile (Table 2), and the user-defined Pr(D) versus the observed frequency of dropout (Fr(D)) for each contributor per profile (Table 3).

2. GGETIt Algorithm

During the simulation, alleles for each locus are chosen by generating a random number from a uniform distribution between 0 and the sum of the allele frequencies, and assigning the allele that corresponds to that random number. The sum of the allele frequencies may be one or greater than one if a minimum allele frequency is applied, according to the 5/2*N* rule, where *N* is the sampled number of individuals, as suggested by the National Research Council [4-5]. For the amelogenin locus, the first allele chosen is always 'X' since both males and females have one 'X' chromosome; the second allele is chosen based on the process previously described. The profile S_C of each contributor C is a sequence of unordered pair of alleles. This representation allows for a simple model of allele expression without taking into account signal intensity [6]. Thus, we have:

$$S_C = \left(\{A_{C,L,1}, A_{C,L,2}\}\right)_{L=1}^{n}, \tag{1}$$

where, for $i = 1,2$, $A_{C,L,i}$ is the ith allele of contributor C at locus L; and n is the number of loci (16 for the Identifiler® Plus simulation and 22 for the GlobalFiler™ simulation). Using this representation for an individual profile, a mixture $M_{\mathcal{C}}$ created by the set of contributors $\mathcal{C}$ can be expressed using the following equation:

$$M_{\mathcal{C}} = \left(\bigcup_{C \in \mathcal{C}} \{A_{C,L,1}, A_{C,L,2}\}\right)_{L=1}^{n}. \tag{2}$$

Once the alleles are generated, the simulation applies dropout based on the user defined dropout probabilities for each contributor. This is accomplished by Bernoulli trial. A random number

uniformly selected from 0 and 1, $d_{C,L,i}$ is generated for each allele $A_{C,L,i}$. The allele $A_{C,L,i}$ 'drops out' if

$$d_{C,L,i} \leq \Pr(D)_C, \quad (3)$$

where $\Pr(D)_C$ is the user defined dropout probability for contributor *C*. After applying dropout, alleles that have not dropped out remain in black text while red text indicates alleles that have dropped out. To generate the 'observed allele output', alleles in red text and duplicate alleles are filtered. Figure 1 shows a representative locus exhibiting allelic dropout. In this example, the second contributor, Person2, is homozygous for allele 14 at the locus D8S1179. Based on a Pr(D) of 0.4 for the second contributor, if the random number 0.15 was generated for Allele1 for this contributor, this instance of the allele would drop out since 0.15 is less than the defined Pr(D) of 0.4. In order for the 'observed allele output' to show the complete dropout of allele 14, both Allele1 and Allele2 would have to dropout.

A

Profile 1	Person1		Person2	
	Allele1	Allele2	Allele1	Allele2
D8S1179	11	12	14	14

B

Profile 1	Person1		Person2	
	Allele1	Allele2	Allele1	Allele2
D8S1179	11	12	14	14

C

Profile 1	Allele1	Allele2	Allele3
D8S1179	11	12	14

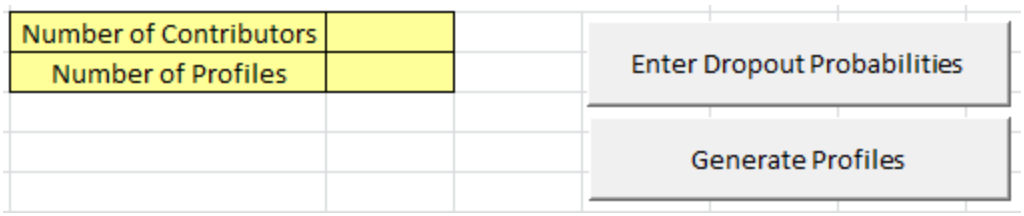
Figure 1. A representative locus where there are two contributors and some allelic dropout. (A) Known genotypes of Person1 and Person2 for the locus D8S1179 prior to the simulation applying dropout. (B) Person2's Allele1 drops out since the random number generated by the simulation (0.15) is less than the Pr(D) the user assigned for this contributor (0.4). The red text indicates the allele has dropped out. (C) The 'observed allele output' for this locus.

3. Requirements and Procedure

1. GGETIt was developed using Visual Basic for Applications in Microsoft® Excel® 2010, version 14.0.7166.5000. GGETIt has been tested on Windows 7 operating system. Enable macros. If needed you may need to change the 'Macro Settings' in Excel, found under the File Tab in Excel Options->Trust Center->Macro Settings.

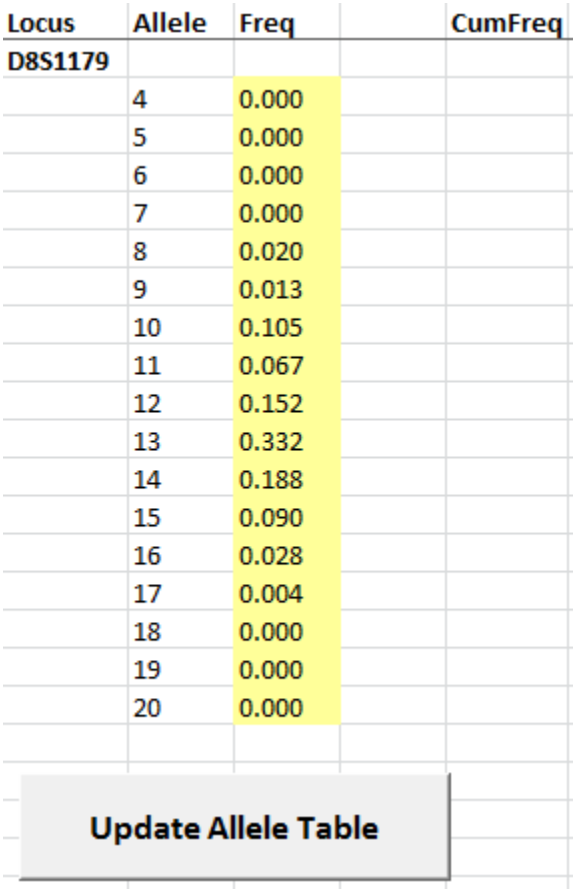
2. Download GGETIt for Identifiler® Plus or GlobalFiler™ from Boston University's DNA Mixture website, http://www.bu.edu/dnamixtures/.

3. Open GGETIt. The following Excel sheet, *Inputs*, will appear:



The allele table in GGETIt on sheet *AlleleTable* contains allele frequencies from the Caucasian population data in the GlobalFiler™ User Guide [2].

4. To use allele frequencies from another source, click on the second sheet of the file, *AlleleTable,* and enter desired allele frequencies into the highlighted cells for each locus. All of the possible alleles for each locus are listed.



5. After entering allele frequencies for all loci, click on the *Update Allele Table* button, located below the allele frequencies for the locus D8S1179 (Identifiler® Plus simulation) or D3S1358 (GlobalFiler™ simulation). This will compute the cumulative allele frequency for each locus. If the allele frequencies do not require updating, skip steps 4 and 5.

Locus	Allele	Freq		CumFreq
D8S1179				
	4	0.000		
	5	0.000		
	6	0.000		
	7	0.000		
	8	0.020	0.000	0.020
	9	0.013	0.020	0.034
	10	0.105	0.034	0.138
	11	0.067	0.138	0.206
	12	0.152	0.206	0.357
	13	0.332	0.357	0.690
	14	0.188	0.690	0.878
	15	0.090	0.878	0.968
	16	0.028	0.968	0.996
	17	0.004	0.996	1.000
	18	0.000		
	19	0.000		
	20	0.000		

Update Allele Table

6. On the *Inputs* sheet, enter the *Number of Contributors* (1-6) and the *Number of Profiles* to generate.

Number of Contributors	
Number of Profiles	

Enter Dropout Probabilities

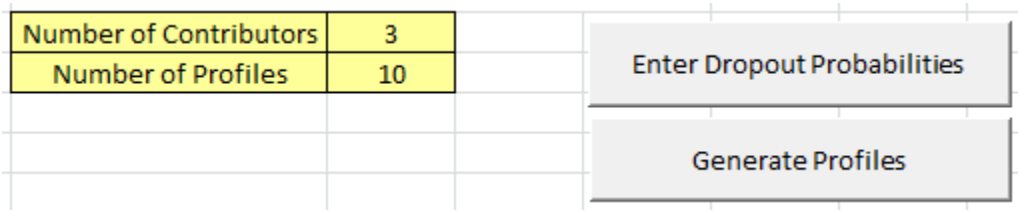
Generate Profiles

Due to the maximum number of rows allowed by Excel, a maximum of 55,188 profiles and 41,943 profiles for the Identifiler® Plus and GlobalFiler™ simulations, respectively, can be generated.

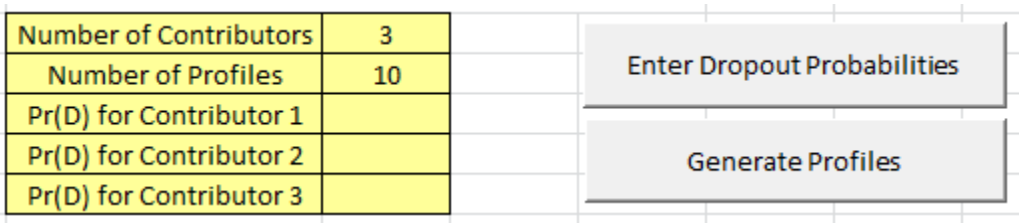
The run time of GGETIt is dependent on the *Number of Profiles* generated. Typical run times for up to 10,000 profiles on a dual-core laptop with Intel® Core™ i5-4200M CPU @ 2.5GHz are detailed in the following table. Generating more than 10,000 profiles has not been tested. Make sure the sleep setting (usually located in the Control Panel of a PC) is turned off. Otherwise, the computer may fall asleep during the simulation. Generating more than 10,000 profiles has not been tested.

Number of Profiles	Approximate Run Time
1	1 second
100	30 seconds
1,000	5 minutes
5,000	23 minutes
10,000	45 minutes

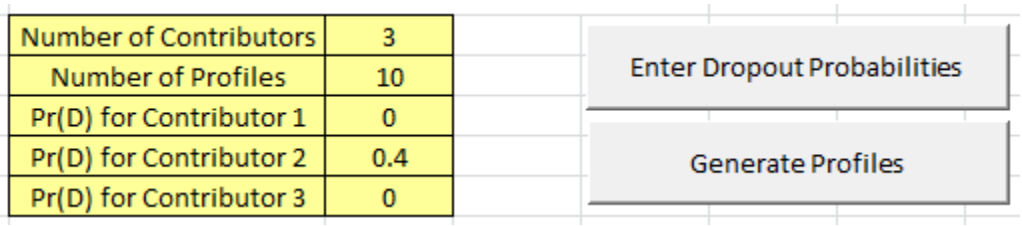
7. Click on the *Enter Dropout Probabilities* button.



8. Enter the *Pr(D)* for each contributor (0-1).



9. Click on the *Generate Profiles* button.



10. After generating the profiles, the 'observed allele output' on the *ObservedProfiles* sheet will be displayed. The *Minimum Number of Contributors* is calculated using MAC, whereby the NOC is determined by counting the number of alleles, dividing by two, and rounding up for each locus. The example below is for the Identifiler® Plus simulation.

Profile 1	***Minimum Number of Contributors = 3***					
	Allele1	Allele2	Allele3	Allele4	Allele5	Allele6
D8S1179	10	13	14	15		
D21S11	28	29	30	31		
D7S820	10					
CSF1PO	9	11	12			
D3S1358	14	15	18			
TH01	6	9	9.3			
D13S317	8	11	12			
D16S539	9	11	12			
D2S1338	18	19	21	24	25	
D19S433	12	13				
vWA	16	17				
TPOX	8	11				
D18S51	14	15	18			
Amel	X	Y				
D5S818	11	12	13			
FGA	19	22	24	25		

11. The 'true profiles', or known genotypes, of each contributor for every profile can be viewed on the *Profiles* sheet, color-coded to delineate the alleles per contributor that did (red) and did not (black) dropout. The example below is for the Identifiler® Plus simulation.

Profile 1	Person1		Person2		Person3	
	Allele1	Allele2	Allele1	Allele2	Allele1	Allele2
D8S1179	13	13	14	10	15	14
D21S11	29	31	28	29	29	30
D7S820	10	10	10	10	10	10
CSF1PO	11	12	10	12	11	9
D3S1358	15	15	14	18	14	15
TH01	6	9.3	9	7	9.3	6
D13S317	11	12	11	12	8	11
D16S539	9	11	11	11	11	12
D2S1338	25	18	21	24	19	24
D19S433	12	13	13	14	13	12
vWA	17	16	16	19	17	16
TPOX	8	8	8	11	8	11
D18S51	15	15	18	14	14	14
Amel	X	Y	X	Y	X	Y
D5S818	11	12	11	13	12	11
FGA	25	24	19	22	22	19

12. The *Results Table1* sheet provides *Table 1 - True vs Observed Number of Contributors* for each profile. Highlighted profiles indicate that the NOC has been underestimated. An *Observed Number of Contributors Summary* is also provided.

TABLE 1 - True vs Observed Number of Contributors

Profile #	True	Observed
1	3	3
2	3	3
3	3	3
4	3	3
5	3	3
6	3	3
7	3	3
8	3	3
9	3	2
10	3	3

Observed Number of Contributors Summary

Number of Contributors	Number of Profiles	Percentage
1	0	0.0%
2	1	10.0%
3	9	90.0%
4	0	0.0%
5	0	0.0%
6	0	0.0%

13. The *Results Table2* sheet provides *Table 2 - Number of Observed Alleles per Locus* for each profile. Highlighted cells indicate that the number of observed alleles falls below the number of expected alleles given that each person in the mixture contributes 1 or 2 alleles. In this example, for a mixture of 3 contributors, 5 or 6 alleles are expected. Less than 5 alleles are observed due to allele sharing or dropout. The example below is for the Identifiler® Plus simulation.

TABLE 2 - Number of Observed Alleles per Locus

Profile #	D8S1179	D21S11	D7S820	CSF1PO	D3S1358	TH01	D13S317	D16S539	D2S1338	D19S433	vWA	TPOX	D18S51	AMEL	D5S818	FGA
1	4	4	1	3	3	3	3	3	5	2	2	2	3	2	3	4
2	3	6	3	4	4	3	4	3	3	2	2	4	4	2	3	4
3	4	5	4	1	4	4	3	3	3	3	3	2	4	2	3	3
4	5	3	5	5	3	3	4	4	5	2	3	2	4	2	2	3
5	4	4	4	3	2	3	3	3	4	3	4	3	5	2	4	4
6	4	4	3	4	3	5	3	3	4	4	3	3	4	2	2	5
7	3	4	3	3	5	3	3	3	4	5	4	3	5	2	3	5
8	5	5	5	3	3	4	4	4	2	4	3	2	2	2	3	4
9	4	3	4	3	3	3	3	3	3	4	4	2	3	2	3	3
10	3	4	3	2	4	4	2	4	3	4	5	2	4	2	2	3

14. The *Results Table3* sheet provides *Table 3 - Probability vs Frequency of Dropout* for each profile. The probability of dropout is defined by the user, and the frequency of dropout is calculated from the 'true profiles'.

Table 3 - Probability vs Frequency of Dropout						
Profile #	Person 1		Person 2		Person 3	
	Pr(D)	Fr(D)	Pr(D)	Fr(D)	Pr(D)	Fr(D)
1	0	0	0.4	0.438	0	0
2	0	0	0.4	0.406	0	0
3	0	0	0.4	0.406	0	0
4	0	0	0.4	0.375	0	0
5	0	0	0.4	0.125	0	0
6	0	0	0.4	0.344	0	0
7	0	0	0.4	0.375	0	0
8	0	0	0.4	0.469	0	0
9	0	0	0.4	0.562	0	0
10	0	0	0.4	0.344	0	0

4. References

1. Applied Biosystems. AmpFℓSTR® Identifiler® Plus PCR Amplification Kit User Guide. Foster City, CA: Applied Biosystems, 2015.
2. Applied Biosystems. GlobalFiler™ PCR Amplification Kit User Guide. Foster City, CA: Applied Biosystems, 2015.
3. Lesson 11 Number of Contributors vs. Number of Alleles Observed. Boston University Biomedical Forensic Sciences DNA Mixtures. Available from: http://www.bu.edu /dnamixtures/18/ (accessed September 24, 2015).
4. Butler JM. Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers. 2nd ed. Burlington, MA: Elsevier Academic Press, 2005.
5. Evaluation of forensic DNA evidence. Washington, DC: National Research Council; 1996.
6. Gordon JS. Characterization of error tradeoffs in human identity comparisons: Determining a complexity threshold for DNA mixture interpretation [Master of Science thesis]. Boston, MA: Boston University School of Medicine, 2012.

5. Acknowledgments

I would like to thank Dr. Catherine Grgicak, for her advice while developing GGETIt. I would also like to thank Jacob Gordon and Desmond Lun for helping me formalize the GGETIt algorithm.