A simulation model of DNA template quality, used in validating a genetic ancestry estimation system for forensic applications

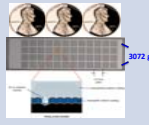
Jason Bryan, BS, Marc Bauchet, Ph.D., Victoria Vance, MS, Dan Hellwig, MFS, Lars Mouritsen, BS
Sorenson Genomics

Abstract

Sorenson Forensics recently released a genetic ancestry estimation test known as Investigative LEAD℠ (Law Enforcement Ancestry DNA). This new test provides a means for law enforcement agencies to identify the genetic ancestry of suspects and/or victims. Software systems used to estimate genetic ancestry may vary based on the type and number of genetic markers, generally Single Nucleotide Polymorphisms (SNPs), observed and statistical algorithms used. The common premise of these systems, however, is to measure the genetic affinity of an individual in relation to representative parental populations. The result is generally reported in terms of relative percentages for each population. Laboratory test systems used to generate SNP data may vary by chemistry and analytical method, which can play a role in the number, quality, and accuracy of the genotypes used to derive ancestry estimations for a given DNA sample. Forensic-type samples often have low DNA copy numbers due to minimal sample amount or degradation and frequently contain inhibitors. These factors can lead to lower recovery rates of targeted loci and allele dropout or stochastic effect. When DNA quality or quantity is compromised, it is important to determine that the accurate genetic ancestry estimations can still be made. Sorenson Forensics' I-LEAD test makes use of the Applied Biosystems TaqMan OpenArray technology to genotype 192 autosomal SNPs. A proprietary algorithm developed at Sorenson is used to create the estimations of ancestry. An ancillary software program (Genotype Degrader) was created to assist in the validation of the algorithm used for the I-LEAD test system. Initially, full genotype profiles for 190 SNPs were generated from DNA samples with known genetic ancestry. Settings were selected within the Genotype Degrader simulation program to randomly introduce specific amounts of locus dropout (up to 50%) and stochastic effect into each given genotype set (up to 30%). The simulation for each parent genotype profile can be run as many times as desired creating innumerable, unique combinations of the genotype set at a desired 'quality-level'. The ancestry estimation of the "degraded" genotypes can then be compared to that of the parent genotype to assess the impact random degradation would have on ancestry estimations. The Genotype Degrader software tool allowed for well-controlled experimentation to demonstrate the accuracy of the I-LEAD test when locus dropout and stochastic effects are observed in genotype data. These simulations greatly reduced the time and expenses for this type of study over actual sample testing in the forensic laboratory.

Background

Data collection: SNPs are tested using the TaqMan® Open Array® Genotyping System from Life Technologies. It uses fluorescence-based polymerase chain reaction (PCR) reagents to provide qualitative detection of targets using post-PCR endpoint analysis. As a modified approach to standard TaqMan® genotyping, this system miniaturizes the reactions down to 33 nanoliters for cost efficiency and high throughput.



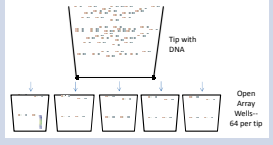
1 well = 1 reaction = 1 locus

3 possible answers at each bi-allelic SNP:

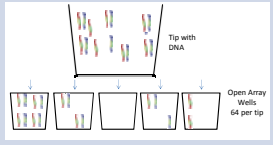
AA, AB, or BB

DNA Template Concentration and Quality Effects on TaqMan® OpenArray® System:

Distribution of DNA from NORMAL CONCENTRATION and/or QUALITY



Distribution of DNA from LOW CONCENTRATION and/or QUALITY



Distribution is random leading to random stochastic dropout of loci and stochastic loss of heterozygosity

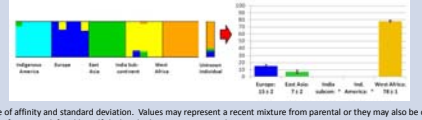
The Sorenson Investigative LEAD℠ Test :

The Sorenson Investigative LEAD℠ Test, a genetic ancestry test based on a selection of 190 informative markers (AIMs) selected to ascertain an individual's genetic ancestry. This ancestry is defined by a reference set of 5 population samples from the International HapMap¹ phase 3 dataset, namely Yoruba (Ibadan, Nigeria) representing West Africa*, Indigenous America† (CEPH Indigenous American Pima, Maya, Karitiana, Surui, and Arawak), Han Chinese (Beijing, China) East Asia*, Europeans* (Utah residents with ancestry from northern and western Europe, USA), and Gujarati Indians (Houston, USA) for India Sub-continent*.

Sorenson Investigative LEAD℠ test calculates a human DNA sample's affinity to those 5 population samples, suggesting the individual's genetic ancestry—a clue that may help determine an individual's physical appearance. We believe this to be an important tool for investigators dealing with DNA samples of unknown or ambiguous origin.

Ancestry Informative Markers (AIMs): A set of Single Nucleotide Polymorphisms (SNPs) selected from large public datasets of nearly 1 million SNPs, chosen for their ability to discriminate among the 5 major worldwide populations and represent every autosomal chromosome.

Data Analysis: Utilizes Principal Component Analysis (PCA) and a proprietary algorithm based on the program *frappe* to calculate affinity levels of an individual DNA sample toward each of the 5 reference populations.



Numeric values indicate degree of affinity and standard deviation. Values may represent a recent mixture from parental or they may also be compared to affinity percentages of other, more specific groups defined by self-declared ethnicities or geographic regions.

Method

The Genotype Degrader program functions with 3 points of input criteria:

1. **Sample genotype set of bi-allelic SNPs**
2. **Target % of stochastic dropped loci**
3. **Target % of stochastic reduced heterozygosity**

The Genotype Degrader program then uses the following formulas to determine the total amount of stochastic quality/quantity to simulate for the input genotype set:

Formula 1

n_t = Number of total genotypes for a test sample
a^1a^2 = Genotype at target SNP locus
x_1 = A, G, C, or T
$x_0 = 0$

$$n_t = \sum_{x_1=a^1a^2} a^1a^2$$

Formula 2

n_d = Number of dropped genotypes to introduce from all genotypes
d = Drop-out % (input)

$$n_d = n_t - dn_t$$

Formula 3

n_s = Number of Stochastic genotype changes to introduce from heterozygous genotypes
s = Stochastic genotype from heterozygous genotype (input)

$$n_s = sn_d$$

Formula 4

Δq = Total change in genotype quality/quantity

$$\Delta q = n_d + n_s$$

Each SNP for "n_s" and "n_s" to be changed is selected at random by the algorithm.
The homozygous allele call change for each SNP of "n_s" is also selected at random by the algorithm.

Samples Tested	Ancestry estimated data set chosen from	Population sets represented	Samples per population set
20	Affinity% attained for 1 population was at or near 100%	5 (5 (ancestry estimation pops))	2
10	Affinity% attained for 2 populations mixed at or near 50% each	10 (10 (2 pop mixed permutations))	1
Total: 20			

A "baseline" target population affinity percentage and standard deviation were first obtained by running the unmodified genotype of each sample 20 times through the ancestry estimation algorithm.

Each sample was then processed through the Genotype Degrader simulation program creating a minimum of 5 unique randomly degraded genotypes at each varied level of stochastic dropped loci and reduced heterozygosity.

Subsequently, each iteration at each permutation was processed through the ancestry estimation algorithm and assigned an affinity percentage to each of the 5 populations.

% Reduced Heterozygosity	% Dropped Loci: 0	5	10	15	20	25	30	35	40	45	50
0	20	10	10	10	10	10	10	10	10	10	10
5	5	x	x	x	x	x	x	x	x	x	x
10	5	x	5	x	5	x	5	x	5	x	5
20	5	x	5	x	5	x	5	x	5	x	5
30	5	x	5	x	5	x	5	x	5	x	5

Total unique genotypes per sample = 196

3,900 "degraded" unique genotypes simulated from 20 original "full" genotypes

Results and Discussion

Samples with Affinity % for 1 population At or Near 100%

Sample data were separated into categories of those where the observed population was at or near 100% or were a relative 50/50 mix of 2 populations.

Observed % affinity of the target population for each permutation set was averaged with a calculated standard deviation and then compared to the "baseline" for that sample.

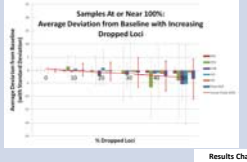
Results Charts 1 and 5 reflect results of increased levels of random dropped loci while heterozygosity percent remained constant.

Results Charts 2 and 6 reflect results of reduced levels of heterozygosity while no artificial locus dropout was applied.

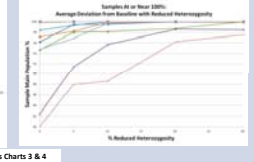
Other validation work, however has shown that dropped loci and loss of heterozygosity typically occur together as a result of reduced DNA quantity/quality, though results may vary depending on the amount of heterozygosity in the sample to begin with.

Results Charts 3 & 4 and 7 & 8 reflect results of combined stochastic simulated dropout and loss of heterozygosity. Charts were separated into two, one for the Average Deviation from the base line and the other for Average Standard Deviation at each permutation set, for formatting reasons.

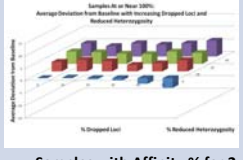
Results Chart 1

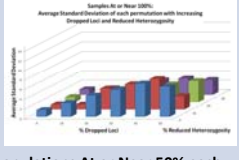


Results Chart 2



Results Charts 3 & 4





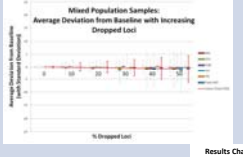
The observed trend for the data set in Results Chart 2 was that the reduced heterozygosity 'forced' population affinity to 100% for all samples tested.

When only dropped loci are factored, the observed trend of samples at or near 100% is to decrease in % affinity with the potential for greater variability in the final call. The % affinity lost from the major population may be shifted to 1 or more of the other 4 reference populations.

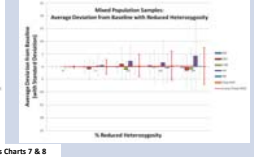
When both factors (dropout and LOH) are included, the observation for samples at or near 100% is that the loss of heterozygosity overpowers the locus dropout and generally drives the population to 100%. Therefore, the variability decreases slightly as the % affinity approaches 100%.

Samples with Affinity % for 2 populations At or Near 50% each

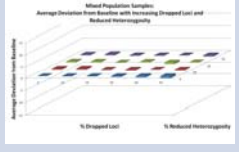
Results Chart 5

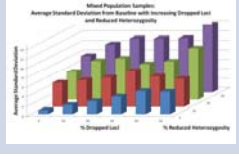


Results Chart 6



Results Charts 7 & 8





When only reduced heterozygosity is factored, the observed trend of mixed samples is variable in % affinity, sometimes up, sometimes down, resulting in an average % affinity similar to the original baseline; however with increased observed variability.

When both factors (dropout and LOH)are included, the observation for mixed samples is that the average % affinity observed at each permutation remains very consistent; however, the effects from both dropped loci and loss of heterozygosity compound resulting in the potential for greater variability.

Citations

*The reference population data sets used in calculations was taken from individuals that are represented in the HapMap 3 project (http://hapmap.ncbi.nlm.nih.gov/). HapMap sample and continental designations found on this report are associated in the following way: European (CEU - Utah , USA residents with Northern and Western Europe ancestry from the CEPH collection and TSI- Toscana in Italia); Asia (CHB – Han Chinese in Beijing, China; CHD –Denver, USA residents with Han Chinese ancestry from the CEPH collection; JPT-Japanese in Tokyo, Japan); India Subcontinent (GIH - Gujarati Indians in Houston, Texas, USA); Africa (YRI – Yoruba tribe in Ibadan, Nigeria and LWK –Luhya in Webuye, Kenya).

†The reference population data sets used in calculations was taken from individuals that are represented in the Human Genome Diversity Project (HGDP). Samples representing herein the "Indigenous Americas" population are from the following HGDP populations: Colombian (Arawak), Karitiana, Maya, Pima, and Surui. Details on the collections see H. Cann et al. Science 296:261-262 (2002) A human genome diversity cell line panel, and its sSUpplemental Data; Rosenberg et al. Science 298: 2381-2385 (2002); and Rosenberg et al. PLoS Genetics 1:660-671 (2005).

Conclusion

The Genotype Degrader program was created to assist in the validation of the I-LEAD Test System. By randomly introducing specific amounts of stochastic locus dropout (up to 50%) and loss of heterozygosity (up to 30%) into each given genotype set, we were able to observe the effects of simulated degradation on the final % affinity values for a given sample. For n=3900 simulations we have observed a well maintained average of <5% deviation from baseline % affinity value. The overall variability and confidence in the reported affinity percentage however, increases depending on total number of dropped loci and loss of heterozygosity which can be variable depending on population(s) being observed. According to the data presented here, the I-LEAD test maintains the ability to generate usable, informative data, even when suboptimal DNA concentration and quality are suspected, so long as the variability and confidence in the final % affinity value is reflected on the final report.

The Genotype Degrader program was used in this study with 190 total bi-allelic SNP data points, however, the program has the ability to be set up to do artificial degradation on much larger SNP genotype sets (on the order of hundreds of thousands). The simulation for each parent genotype profile can be run as many times as desired creating innumerable, unique combinations of the genotype set at a desired 'quality-level'. The Genotype Degrader program allowed for well-controlled experimentation over thousands of simulations to demonstrate the accuracy of the I-LEAD test when stochastic locus dropout and loss of heterozygosity effects are observed in genotype data.

These simulations greatly increased the level of control for this type of study and reduced the time and expenses over actual sample testing in the forensic laboratory. With only the reagent expense of a few samples, a database set of nearly 4,000 unique genotypes was generated and tested.

Sorenson Forensics
2495 S. West Temple
Salt Lake City, UT 84115
SorensonForensics.com