Estimating Genetic Ancestry

Using the Investigative-LEAD℠ *(Law Enforcement Ancestry DNA)* Test

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

ABSTRACT

The utilization of many worldwide DNA databases is an essential tool in modern criminal investigations. Unfortunately, when an evidentiary DNA profile does not provide a viable suspect subsequent to a database search, the investigator may be left with little forensic direction. To assist in these critical situations, Sorenson Forensics introduces Investigative LEAD ; a single nucleotide polymorphism (SNP) based DNA test designed to estimate genetic ancestry against a model of 5 genetically distinct, putative parental populations. The populations and the reference samples representing them are as follows: Western European (HapMap CEU, Northwest European descent residing in Utah), West Sub-Saharan African (HapMap YRI, Yoruba from Ibadan, Nigeria), East Asian (HapMap CHB from Beijing, China), Indigenous American (Compilation of samples identified as being from populations indigenous to North, Central, and South America including Maya, Pima, Karitiana, Surui, and Arawak descent), and the India Subcontinent (HapMap GIH, Gujarati Indian descent residing in Houston, TX). Our method uses 190 SNP Ancestry Informative Markers (AIMs) chosen from their scored ability to specifically differentiate between the 5 reference populations using Principal Component Analysis (PCA) as the comparative analysis tool and includes some markers identified as informative in previous genetic ancestry estimation publications. Using the program FRAPPE and uniquely designed algorithms, the method compares an unknown individual sample to at least a hundred randomly selected subsets of individuals from the reference populations. Background interference is calculated simultaneously and is used to estimate confidence intervals based on a calibration that was effected using thousands of worldwide individuals. Validation data have shown the Investigative LEAD test is a viable, robust and adequately sensitive test, capable of functioning on a variety of different forensic samples and DNA extract types. We believe this test will provide law enforcement investigators valuable information regarding the genetic ancestry of potential suspects. This test can be a great benefit for solving cold cases and other criminal investigations.

Sorenson Forensics®

2495 S. West Temple, Salt Lake City, Utah 84115

SorensonForensics.com

Materials and Methods

We present Sorenson LEAD℠, a genetic ancestry test based on a set of ancestry informative markers (AIMs) selected to ascertain an individual's genetic ancestry. This ancestry is mapped by a reference set of five population samples: four from the International HapMap3 1 dataset, namely Yoruba (Ibadan, Nigeria) representing West Africa, Han Chinese (Beijing, China) for East Asia, Europeans (Utah residents with ancestry from northern and western Europe, USA), Gujarati Indians (Houston, USA) for the Indian Sub-continent, and one from the CEPH-HGDP (Pima, Maya, Karitiana, Surui, and Arawak) representing Indigenous Americans. We selected the SNP AIMs that had the most influence in PCA patterns in the ~1 million SNPs datasets. *(refs HM3 , Herraez Bauchet et al)*

Sorenson Investigative LEADSM 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.

SNPs are cost effective, quick to genotype and cover the entire genome.

Data collection: SNPs are tested using the ***TaqMan® Open Array® Genotyping System*** from Life Technologies *(Figure 1)*. The method 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 *(Figure 2)*.

Figure 1. TaqMan® Open Array® Genotyping System

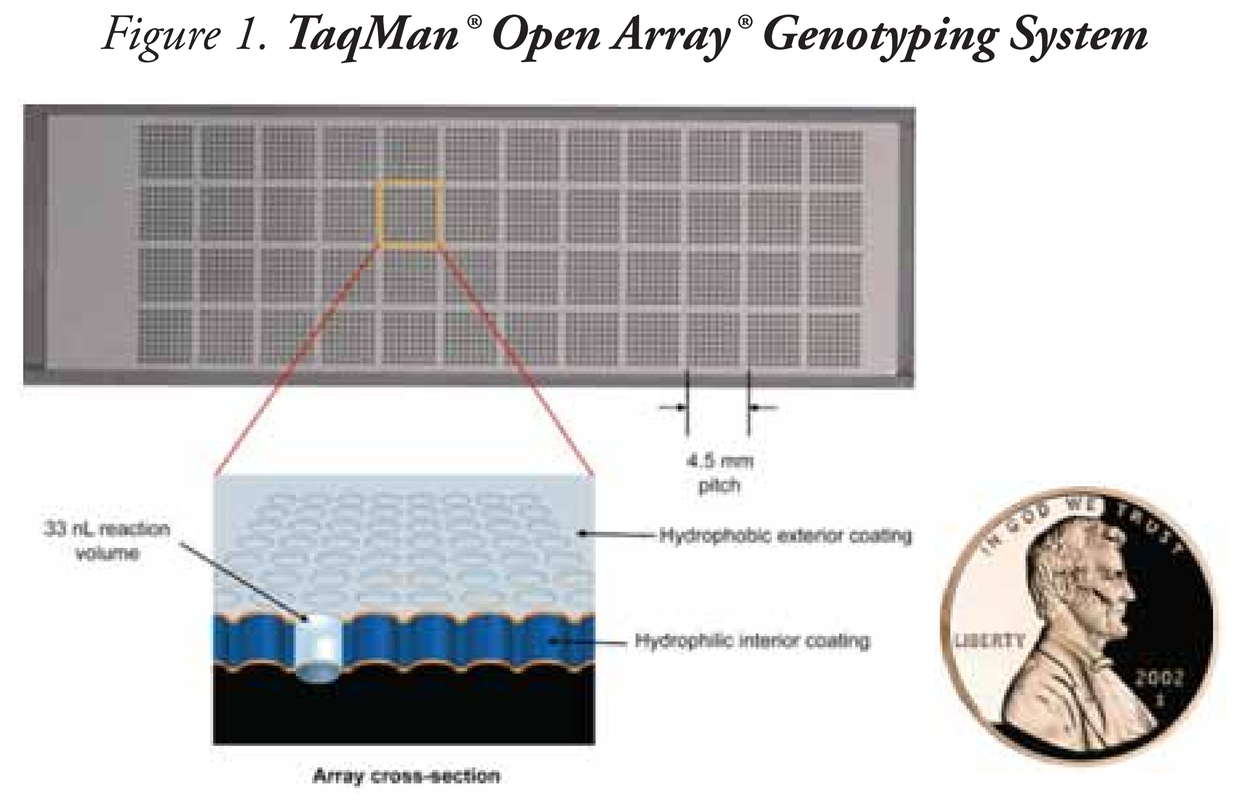

Figure 2. TaqMan® Open Array® Genotyping array card

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 *(Figure 3)*.

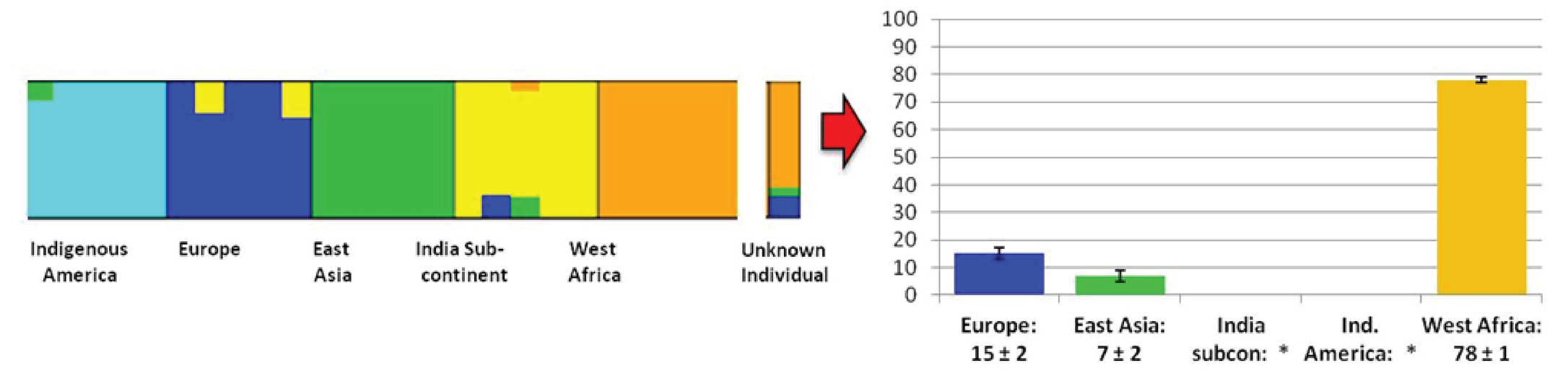

Figure 3. Frappe plot and resulting genetic affinity values +/- SD as calculated by I-LEAD Test System.

Numeric values indicate degree of affinity and standard deviation. Values may represent a recent mixture from parental populations a shown in Figure 4 for an individual with known West African and West European ancestry. Genetic affinity values may also be compared to affinity percentages of other, more specific groups defined by self-declared ethnicities or geographic regions as shown in figure 5. mtDNA and/or Ycs data could be used to supplement results and potentially add specificity.

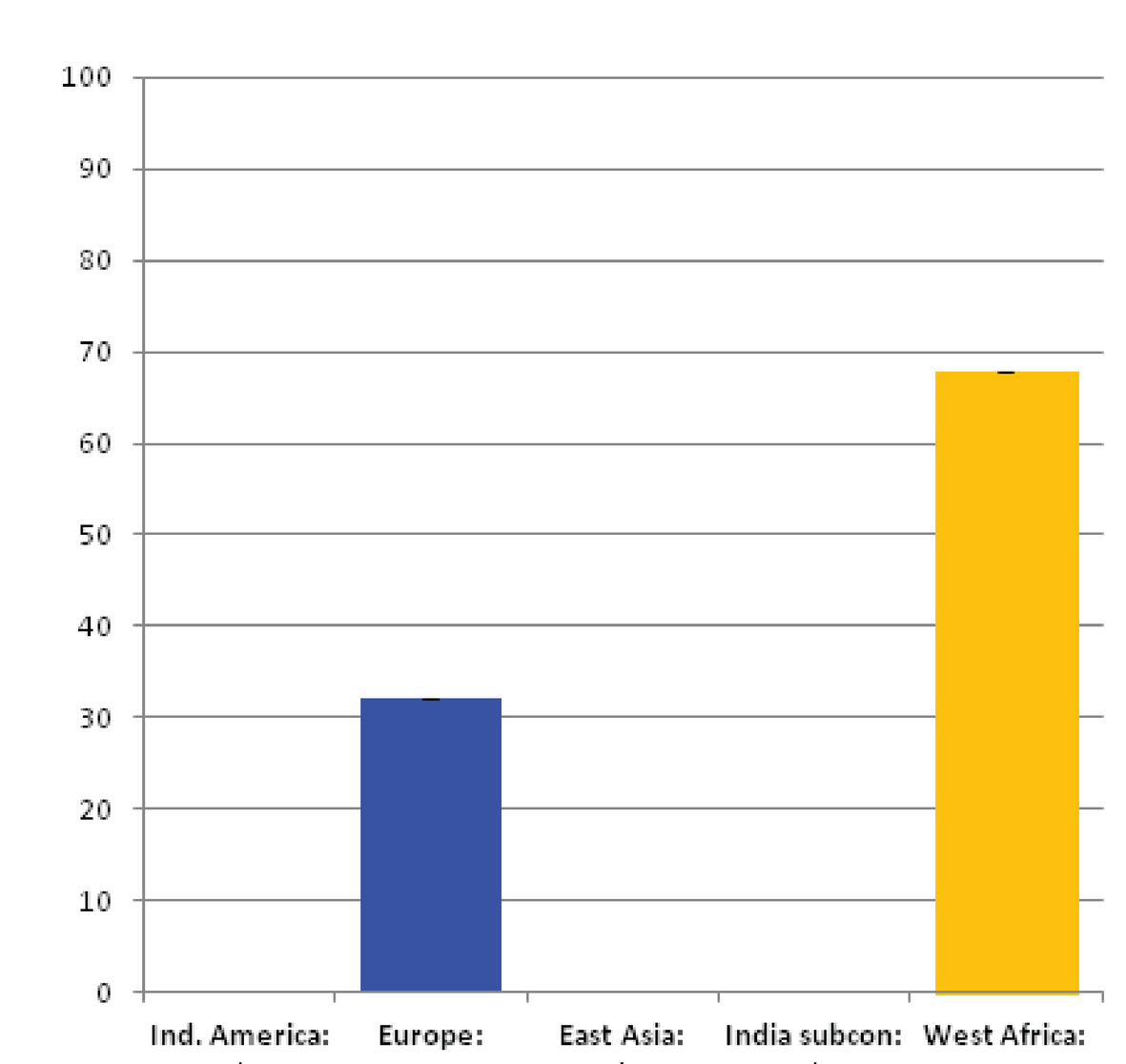

Figure 4. Genetic affinities for an individual with known ancestry from West Africa and Western Europe.

Symbol	Population	n	Symbol	Population	n	Symbol	Population	n
ADY	Adygei	5	GUY	Guyana	2	NAS	Nasioi	5
ASW	Afro American (SW)	44	HAN	Han	5	NAX	Naxi	4
BAL	Balochi	5	HAW	Hawaii	1	ORC	Orcadian	5
BAN	Bantu	5	HAZ	Hazara	5	PAL	Palestinian	5
BAS	Basque	5	HEZ	Hezhen	4	PAP	Papuan	5
BED	Bedouin	5	HUN	Hungary	5	PIM	Pima	5
BER	Bergamo	4	IRA	Iranian	5	PTR	Puerto Rico	5
BIA	BiakaPygmy	4	JAP	Japanese	20	QUE	Quechua	5
BRA	Brahui	5	KAL	Kalash	5	RUS	Russian	5
BUR	Burusho	5	KAR	Karitiana	5	SAN	San	4
CAM	Cambodia	5	LAH	Lahu	5	SAR	Sardinian	5
CEU	European (Utah)	45	LWK	Luhya (Kenya)	25	SHE	She	5
CHB	Chinese (Beijing)	45	MAK	Makrani	5	SUR	Surui	5
COL	Colombian (Arawak)	5	MAN	Mandenka	5	TSI	Tuscan (Italy)	20
DAI	Dai	5	MAO	Maori	1	TU	Tu	5
DAU	Daur	5	MAY	Maya	5	TUJ	Tujia	5
DRU	Druze	5	MBU	MbutiPygmy	5	UYG	Uygur	5
EGY	Egypt	1	MEX	Mexican (LA)	42	VEN	Venezuela	5
FCN	French Canadian	15	MIA	Miaozu	5	XIB	Xibo	5
FIL	Filipino	1	MKK	Masai (Kenya)	50	YAK	Yakut	5
FRE	French	5	MON	Mongola	5	YRI	Yoruba (Nigeria)	45
GIN	Gujarati (Houston)	33	MOR	Morroco	5			
GIR	Gujarati (Houston)	51	MOZ	Mozabite	5			

Table 1. Number of individuals tested on I-LEAD from key worldwide populations.

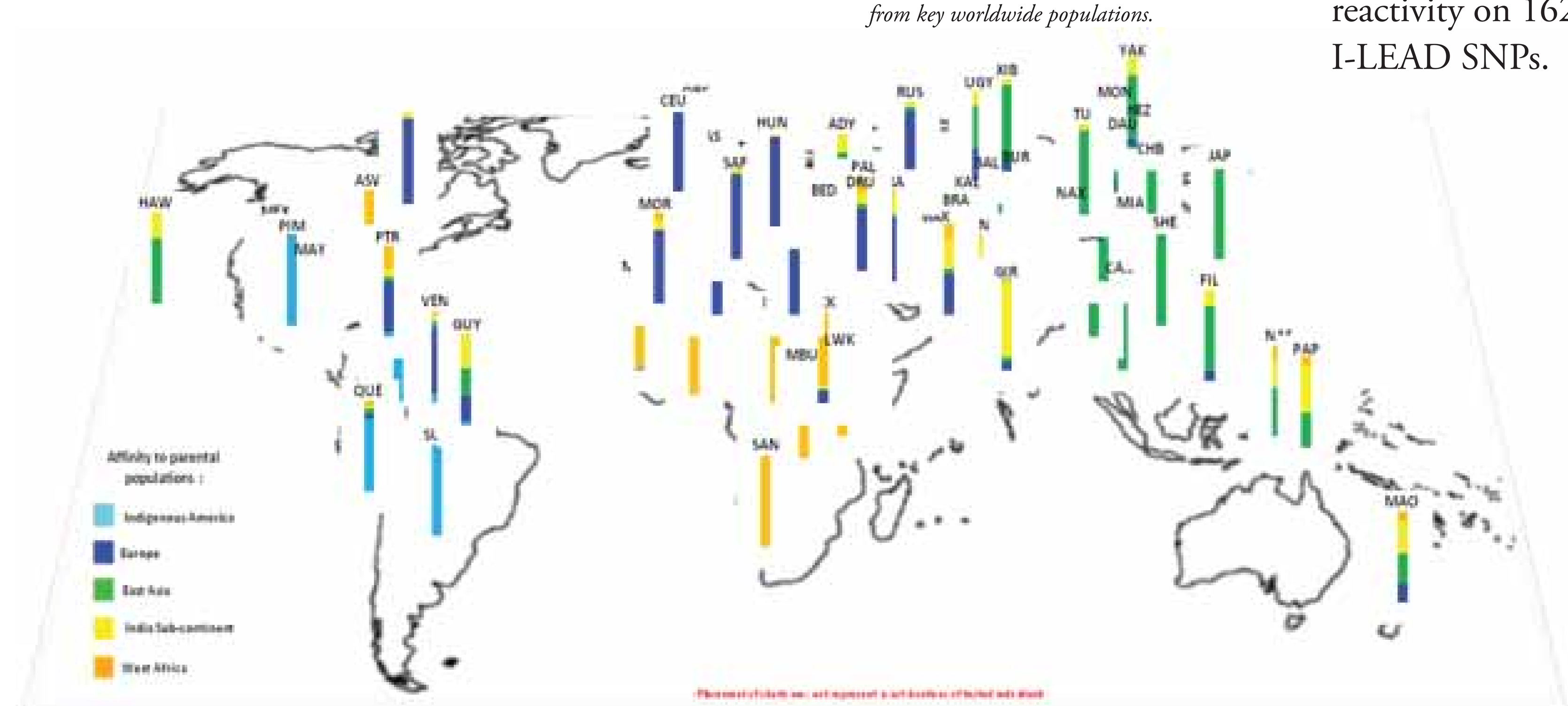

Figure 5. Frappe from worldwide populations as calculated using the I-LEAD Test System.

Interfering Substances

Samples containing varying concentrations of Heme, Indigo, Humic acid, Ethanol, Tannic acid and Calcium Chloride were extracted and tested using a set of 64 quality control SNPs. The purpose of this study was to measure the effect of various interfering substances in the Open Array TaqMan genotyping system. The results are shown in Table 2.

Sample Description	Total DNA (ng)	Loci	Called	%	
ST2001644_Heme_10uM	1.98	52	52	100.00	
ST2001651_Heme_25uM	5.675	62	16	25.81	
ST2001643_Heme_50uM	2.925	62	0	0.00	
ST2001646_Indigo_1uM	3.725	62	60	96.77	
ST2001649_Indigo_4uM	3.025	62	34	54.84	
Indigo Dye* (Replaced ST2001647_Humic_50)	0.9625	62	62	100.00	*ALL FAM results
ST2001704_Humic_10	1.4	62	60	96.77	
ST2001720_Humic_25	1.97	62	36	58.06	
ST2001648_Humic_50	2.7	62	58	93.55	
ST2001652_EtOH_1**	1.9	3	3	100.00	**ALL INV results
ST2001653_EtOH_2**	1.8	0	0	0.00	**ALL INV results
ST2001654_EtOH_3**	1.5	0	0	0.00	**ALL INV results
ST2001706_Tannic_25	1.67	62	9	14.52	
ST2001719_Tannic_75	0.43	62	9	14.52	
ST2001705_Tannic_100	0.24	62	4	6.45	
ST2001708_CaCl2_10uM	1.1	62	0	0.00	
ST2001707_CaCl2_25uM	0.89	47	1	2.13	
ST2001718_CaCl2_50uM	0.24	62	0	0.00	

Table 2. Influence of Interfering Substances on SNP analysis.

Species Specificity

Samples from 5 different model organisms (Yeast, E. coli, Dog, Cow and Chimpanzee) were tested using 189 SNPs from the I-LEAD SNP array. Only Chimpanzee showed some cross reactivity on 162/189 (85.7%) of I-LEAD SNPs. See Table 3.

Species	Callable Genotypes
Yeast	0/189
E. Coli	0/189
Dog	1/189
Cow	1/189
Chimp	162/189

Table 3. SNP detection for non-human species using the I-LEAD Test System.

DNA Extraction Method Comparison

Two different forensic type samples were extracted using six different DNA extraction/purification methods. When DNA was in an appropriate concentration range above 0.4 ng/µL, the overall genotype call rate for all methods was 97.14% (See Table 4). Figure 6 shows the robustness of genotype call rates as the DNA concentration decreases. In general, the I-LEAD Test System is very robust for both forensic samples and standard swab samples down to at least 7.5 ng of total input DNA. The overall genetic affinity calls for samples shown in Figure 6 also proved to be robust down to 1.5 ng of total input DNA (data not shown). The recommended lower limit of input DNA for forensic samples is 3 ng. These data suggest that the I-LEAD method is a robust genotyping platform for DNA extracted using a broad range of extraction methods.

DNA Extraction Method	% Called
Chelex	96.77
DNA IQ	95.16
Organic w Microcon	95.16
Organic w Nucleospin	100.00
Prepfiler	97.35
Qiagen	98.39

Table 4. Call rate by DNA extraction method.

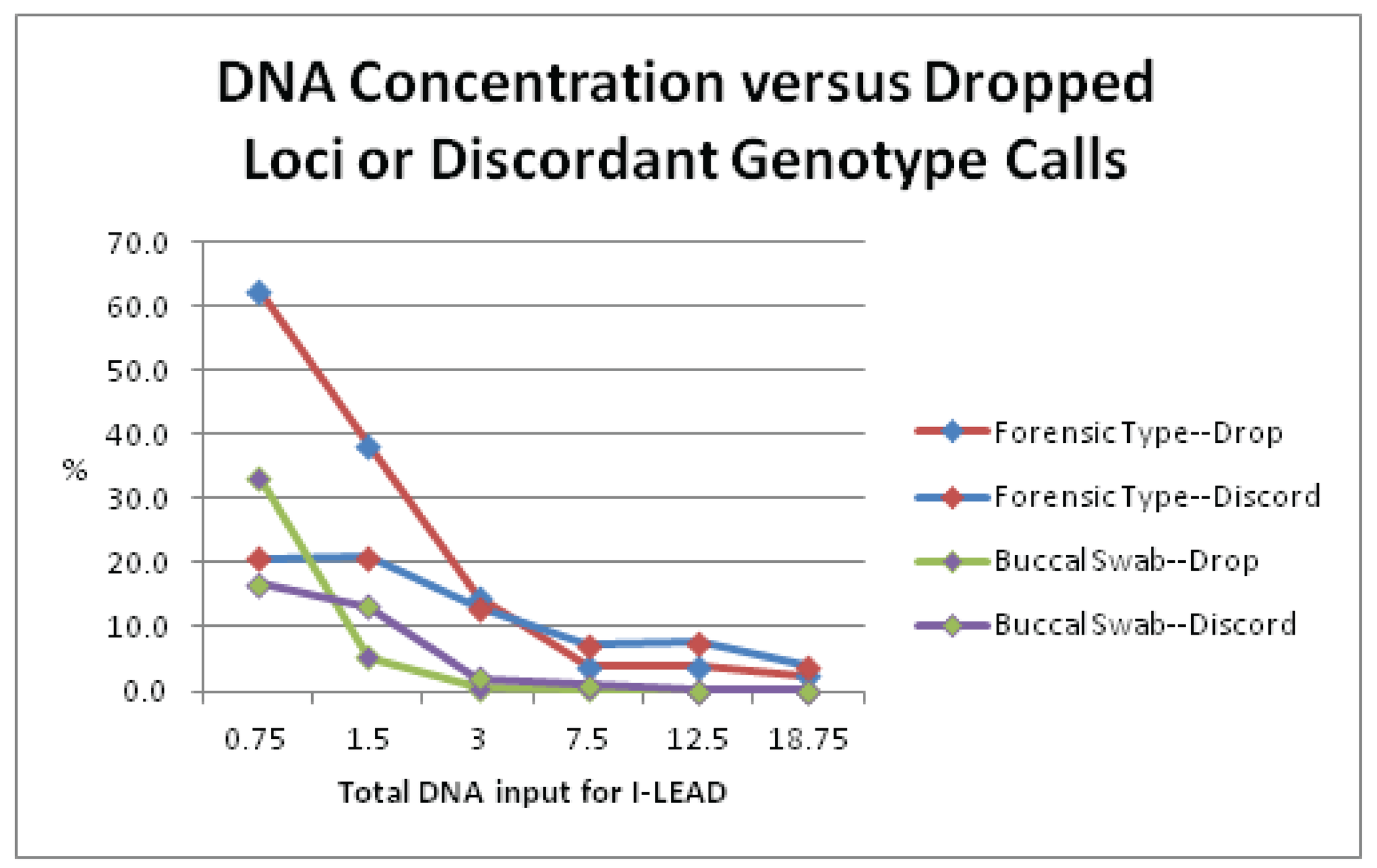

Figure 6. DNA concentration versus Genotype call rate.

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).