First All-in-One Diagnostic Tool for DNA Intelligence: Genome-wide Inference of Bio-Geographic Ancestry, Appearance, Relatedness and Sex with the Identitas v1 Forensic Chip

Aruna T. Bansal[1], Brendan Keating[2], Jonathan Millman[3], Jonathan Newman[3], Kenneth Kidd[4], Bruce Budowle[5], Arthur Eisenberg[5], Joseph Donfack[6], Paolo Gasparini[7], Zoran Budimlija[8], Laurence Rubin[1], Nicholas G. Martin[9], Timothy D. Spector[10], and Manfred Kayser[11] on behalf of the International Visible Trait Genetics (VisiGen) Consortium

[1]Identitas Inc, 1115 Broadway, 12th Floor, New York, NY10010, USA
[2]The University of Pennsylvania, Office 1016, Abramson Building, 3615 Civic Center Bvld, Philadelphia, PA 19104-4399, USA
[3]Centre of Forensic Sciences, 25 Grosvenor Street, Toronto, ON M7A 2G8, Canada
[4]Yale University School of Medicine, PO Box 208005, New Haven, Connecticut 06520-8005, USA
[5]Institute of Applied Genetics, Department of Forensic and Investigative Genetics, University North Texas Health Science Center, 3500 Camp Bowie Blvd, Fort Worth, Texas 76107, USA
[6]Federal Bureau of Investigation, Laboratory Division, 2501 Investigation Parkway, Quantico, VA-22135, USA
[7]Institute for Maternal and Child Health, IRCCS Burlo Garofolo, University of Trieste, Piazzale Europa1, 34127 Trieste, Italy
[8] New York City Office of Chief Medical Examiner, 421 East 26th Street, New York 10016, USA
[9]Queensland Institute of Medical Research, Locked Bag 2000, Royal Brisbane Hospital, Herston, Queensland 4029, Australia
[10]Department of Twin Research, King's College London, St Thomas' Hospital, Westminster Bridge Road, London SE1 7EH, UK
[11]Department of Forensic Molecular Biology, Erasmus MC University Medical Center Rotterdam, PO Box 2040, 3000 CA Rotterdam, The Netherlands

Corresponding author: AT Bansal: bansal at identitascorp dot com; (+44) 1223 421 662

Abstract

When a DNA sample cannot be directly matched to a known, previously genotyped individual, alternate avenues of investigation can be costly and time-consuming. We report on an international collaborative effort to determine the nature and quality of inferences that may be drawn from no-match forensic DNA samples by using a novel SNP marker panel. A total of 3196 DNA samples were genotyped across a highly-targeted panel of genetic markers for bio-geographic ancestry, phenotype and kinship. Trait predictions were performed, blind to the true characteristics of the individuals. Correct predictions were obtained across a range of sample types, from DNA amounts as low as 1.75ng. DNA quality, as opposed to concentration, was a more important indicator of prediction success: highly degraded or highly contaminated DNA tended to fail quality checks and led to fewer predictions. A total of 95% of samples passed quality checks and for those, prediction accuracy was high. Estimates of prediction-accuracy are provided for gender, exact match, 1st-3rd degree relatedness, bio-geographic ancestry, hair color and eye color. These predictions, taken together with matrilineal and patrilineal bio-geographic ancestry, inferred from mitochondrial and Y-chromosome haplogroups respectively, build up a visual profile of the person to whom the sample belongs. The new marker-panel has a wide range of applications to crime-scene, missing persons and mass disaster investigations, as well as matters of homeland (national) security.

Introduction

There have been, and likely will continue to be, forensic cases where the evidentiary DNA profile does not directly match that of a known individual, or any reference sample profile contained within a national DNA database. In addition, current forensic DNA profiling has provided, and likely will continue to provide, little or no information in a number of missing person cases, including mass disaster identification, where scant information is available on the putative identity of the remains found. Traditional policing places heavy reliance on human eyewitnesses to enable investigators to identify suspects. While eyewitness reports have been shown to be helpful, they are highly error-prone [1-2], and consequently a number of people convicted on the basis of eyewitness identification evidence, have been exonerated through forensic DNA testing [2].

In an international, industry-academic collaboration, the International Visible Trait Genetic (VisiGen) Consortium, the Identitas v1 Forensic Chip was developed. The chip, based on well-established Illumina Infinium technology, allows simultaneous genotyping of 192,658 autosomal SNPs of genome-wide distribution, 3,012 Y-chromosomal, 5075 X/ XY-chromosomal, and 428 mitochondrial SNPs. Herein, the first performance study of the Identitas v1 Forensic Chip is reported, based both on data established by consortium members, and data from governmental forensic labs in the USA and Canada. A total of 3196 DNA samples collected from around the world were analyzed. Many of them had recorded sex, continental ancestry, and eye and hair color information, and in some cases, details of relatedness. The DNA samples were of varying quality and quantity as a result of titration and degradation experiments, and the establishment of mock case-work samples.

Genotype quality was assessed, and predictions of sex, bio-geographic ancestry, hair color, eye color and kinship were derived and compared with study-recorded trait data, where available. This study provides the first insights into the performance and feasibility of the Identitas v1 Forensic Chip, the first all-in-one diagnostic tool dedicated to DNA intelligence.

Methods

DNA Samples and Site Reported Data

A total of 3196 DNA samples were studied. A subset of 2780 individuals had site-reported ancestry, and for the purposes of obtaining accuracy estimates, individuals were categorized into 5 major bio-geographic groups, plus another category that included for example, West Asians and individuals from Oceania, for whom HapMap v3 reference samples were not available. The breakdown of site-reported ancestry, where available, was as follows (count in parentheses). 1880 individuals were categorized as European: 240 individuals were categorized as East Asian; 176 individuals were categorized as African descent; 123 individuals were categorized as South American descent; 31 individuals were categorized as South Asian descent; 330 were categorized outside of the five groups. The majority of the aforementioned DNA samples came from TwinsUK and the QIMR Twin Registry studies, as well as from the Yale collection as described in detail elsewhere [3-5]. For those samples, DNA was derived from whole blood either directly or from blood-derived lymphoblastoid cell lines.

Additionally, a subset of 171 DNA samples was derived from more forensically-relevant sources which included (counts in parentheses): hair (2); buccal swab (97); blood swab (9); semen (2); vaginal swab (3); saliva (3); mucus (1); gum (3); drink container swab (8); cigarette butt (10); chap-stick swab (1); swab of tape ends(1); saliva/blood mixture (1); vaginal swab/semen mixtures (30). Two groups contributed sexual-assault type samples. Twenty-nine samples, consisting of either vaginal or buccal samples from female donors (n=8), were spiked with varying amounts of semen from male donors (n=3) and subjected to a standard differential extraction procedure. Results derived using Plexor HY (Promega), a dual autosomal and male-specific quantification assay, indicated that the percentage of male DNA ranged from 100% to none. One group submitted an additional mixture made up of vaginal swab DNA plus semen.

Sensitivity samples were derived by serial dilution of five reference sample extracts of 500ng to 50pg total DNA, yielding total DNA concentrations for each sample of 25 ng/µl, 2.5 ng/µl, 0.25 ng/µl, 0.025 ng/µl and 0.0025 ng/µl. Measurements were taken using Plexor HY (Promega). Degraded samples were experimentally derived from four pre-extracted DNA samples using (a) titrated DNase treatments and (b) by subjecting samples to ultra-violet (UV) light time-courses.

Prior to chip genotyping, DNA samples collated from the different collaborator sites were re-quantified using a picogreen-based assay (Life Technologies). Genotyping followed the standard Illumina Infinium iSelect protocol (www.illumina.com). Fluorescence intensities were detected by the Illumina iScan, and analyzed using Illumina's Beadstudio software. The reaction volumes were 2µl for quality-checking and 5µl for genotyping; additional volume was required to allow for pipetting. In the data received from Illumina, samples were categorised as either 'passed' or 'failed' using the standard control metrics from Illumina, with failure assigned to samples with more than 10% missing genotypes calls.

Statistical Analyses

Sex was inferred on the basis of X-chromosome heterozygosity and the presence or absence of Y-chromosome genotypes. Bi-parental bio-geographic ancestry predictions were conducted using autosomal SNPs exhibiting low linkage disequilibrium. Patrilineal ancestry was derived from 484 Y-chromosmal SNPs and matrilineal ancestry was derived using 280 mitochondrial SNPs, for which phylogenetic and geographic origin information were available [6-8]. Model-based clustering for biogeographic ancestry was conducted using, as a reference, data from HapMap version 3 [9-11]. Hair color and eye color were predicted using multinomial logistic regression models of predictive SNPs as described elsewhere [12,13] except that the following four hair colour predicting DNA variants from the *MC1R* gene could not be implemented on the chip: N29insA (INDEL), Y152OCH, rs1805007 and rs1805009; the hair color prediction model used was adjusted accordingly. Degrees of relatedness were inferred for each pair by calculation of the proportion of the genome shared identical by state (IBS) based upon 192,576 autosomal markers with minor allele frequencies greater than 1%, and less than 5% missing genotypes [14].

Results

Sensitivity

Serial dilutions were conducted by one group for 5 DNA samples extracted from buccal samples of 5 individuals of European bio-geographic origin. Each was diluted to contain 175ng, 17.5ng, 1.75ng 0.175ng or 0.0175ng in the 7µl reaction volume for chip genotyping. For the lowest concentration of 0.0175ng reaction DNA, all five samples failed platform QC with overall genotype call rate <90%, none had the full complement of markers for hair and eye color prediction, and only 1 provided an accurate prediction of bio-geographic ancestry (quantitative method; see later). At the next level of DNA concentration, 0.175ng reaction DNA, one sample passed platform QC, one had the full complement of markers for hair and eye color prediction and three provided accurate predictions of bio-geographic ancestry. At the next level, 1.75ng reaction DNA, three passed platform QC, three had the full complement of markers for hair and eye color prediction, and all 5 provided accurate predictions of bio-geographic ancestry. At higher concentrations, amounting to 17.5ng and 175ng total DNA, 9/10 passed platform QC, 8/10 had the full complement of markers for hair and eye color prediction, and all 10 provided accurate predictions of bio-geographic ancestry.

Despite the preliminary character of the sensitivity testing performed here with small sample-sizes, these results demonstrate that bio-geographic ancestry may be accurately predicted from as little as 1.75ng DNA or even, in some cases, as little as 0.175ng DNA.

Degradation

Twenty-four samples derived from four initial DNA samples were subjected to severe ultra-violet degradation. Upon genotyping, only three passed platform quality checks. Seven samples however, provided accurate predictions of bio-geographic ancestry. The same twenty-four samples were subjected to less severe, enzymatic degradation. Upon genotyping, 19 samples passed platform quality checks however, all 24 enzymatically degraded samples led to accurate estimates of bio-geographic ancestry.

Sexual Assault-type Samples

A total of 30 multi-source DNA samples from simulated sexual assault material extracted after differential lysis were examined. A total of 19/30 samples passed quality checks; however, upon un-blinding at source, it emerged that only 13 of them contained male DNA. For all 13 samples, the Y-chromosome haplogroup was obtained, the geographic origin of which was found to be consistent with site-reported bio-geographic ancestry.

Platform Quality Check

Across the whole study, a total of 3034 (95%) samples passed platform quality checks with overall genotype call rates >90%, while 162 samples (5%) failed this threshold. In the following sections results are presented of the analysis of the 3034 DNA samples that passed quality control checks.

Inference of Sex

A two-pronged approach was taken. Y-chromosome haplogroups, derived from known non-recombining male-specific SNPs were obtained for 1,114 DNA samples, indicating that they were derived from males. In addition, X-chromosome heterozygosity was determined for all samples on the basis of 5,066 X-chromosome specific markers.

Twelve conflicts, between chip-predicted and site-reported sex information were obtained in the 1588 samples with site-reported sex information available. Four samples were predicted to be derived from males on the basis of both identified Y-haplogroup and low X-chromosome heterozygosity, but were un-blinded as being derived from females from the records. Seven samples were predicted to be derived from females on the basis of no inferable Y-haplogroup and high X-chromosome heterozygosity, but were un-blinded as being derived from males from records. Further non-genetic sex data could not be obtained for these eleven individuals. A plausible explanation however, is that DNA mix-ups at some stage may have occurred for these samples, which would correspond to a male-female sample mix-up rate of 0.69% in our study.

Inference of Continental Bio-geographic Ancestry

Bio-geographic ancestry was investigated in three complementary ways: i) determination of mitochondrial haplogroups, ii) determination of Y-chromosome haplogroups and iii) the analysis of genome-wide autosomal genetic data. A quantitative, model-based clustering approach was applied to the autosomal data, which led to the assignment of a probability for each of the five reference groups namely, Africa, Europe, East Asia, South Asia, or South America, based upon HapMap 3 reference data [10]. By this quantitative approach, samples were assigned to a single continental ancestry group whenever the probability for that group was greater than 0.70. When the maximum probability for any single continental group was ≤0.70, the sample was assigned to 'multiple groups'. In this way, 89% of samples were assigned to a single continental or sub-continental group; the remainder were assigned to multiple defined groups. Table 1 shows the results of the analyses. It is noted that, without exception, all of the 17/145 individuals predicted to be of African origin but with conflicting site report, carried mitochondrial and/or Y chromosomal haplogroups of African origin.

Prediction	N	% Correct by Site Report
Africa	145	88%
Europe	1877	93%
East Asia	233	94%
South Asia	24	96%
South America	107	98%

Table 1: Prediction accuracy for individuals with site-reported ancestry

Our approach was insightful also in terms of genetically classifying individuals of mixed continental ancestry. Taking a single example, Figure 1 shows the quantitative assessment of an individual whose father was of European descent and whose mother was of Chinese descent, according to the records. The almost equal proportions of East Asian and European DNA were accurately captured by the method. Furthermore the determination of the Y-chromosome haplogroup as G-L30, which is observed in parts of Europe, Asia and Africa, and the mitochondrial haplogroup D5, which is observed in Eastern Eurasia, indicated that the paternal line is European, whilst the maternal line is East Asian, in agreement with record-based ancestry.

Inference of Eye Color and Hair Color

A total of 1136 samples passed platform quality checks, and had both site-reported eye color, as well as a complete genotype profile for the 6 SNPs required. Within this set 70% of predictions of blue eyes and 85% of predictions of brown eyes agreed with site-reported eye color using the p>0.7 threshold recommended previously [15].

A total of 1137 samples passed platform quality checks, and had both site-reported hair color as well as the complete genotype profile of the 18 SNPs required. Using the previously-developed prediction guide [52] 58% of predictions of black/ dark brown hair corresponded to site report of black, dark brown or brown hair; 72% of predictions of brown/ light brown/ dark blonde corresponded to site report of dark brown, brown, light brown or dark blonde hair; 63% of predictions of blonde/ dark blonde corresponded to site report of light brown, dark blonde, or blonde hair; and 48% of predictions of red hair corresponded to site report of red hair.

Inference of Relatedness

The proportion of the genome shared identical by state (IBS) was estimated for all pair-wise combinations of the 3034 samples, using 192,576 markers. For the majority of the samples however, the true relationships were not known/ site-reported. The true relatedness of samples was available from one source, for which 3240 pair-wise comparisons had been performed for 81 samples. In this set, all 27 first degree relative-pairs, ten second degree relative-pairs (4 uncles/ aunts, 5 grandparents and 1 half-sibling), three third degree relative pairs (2 first cousin pairs and 1great-aunt) and 3199 unrelated pairs were correctly identified. One additional pair of individuals was observed to share 9% of the genome IBS, in line with a 4th degree relationship (e.g. first cousin once removed). Site-report for the two was unrelated, but both were of Caribbean origin. It is acknowledged that the prediction of 4th degree relationships is only valid in highly out-bred populations. For populations which are, or have been in the past, genetically isolated, there will be an over-prediction of distant relatives.

Discussion

The Identitas v1 Forensic Chip is the first all-in-one diagnostic tool targeted for DNA intelligence purposes, allowing for massively parallel genome-wide inference of ancestry, appearance, relatedness, and sex. This chip, manufactured by Illumina using their well-established Infinium technology exhibited, in samples that passed quality control, high predictive power for inference of sex, continental bio-geographic ancestry, and familial relatedness up to 3rd degree level.

The greatest value of the Identitas v1 Forensic Chip relative to other tools for DNA intelligence is that analysis of ancestry, appearance, relatedness and sex are combined in a single all-in-one tool. In criminal investigations, in the absence of a match with a reference sample, such insights can dramatically focus downstream investigations. The DNA-based investigative intelligence obtained can be used in conjunction with, or in the absence of human eyewitness information, to lead potentially, to the identification of suspects. The highly accurate inference of relatedness opens further avenues of application, including paternity/ relationship resolution, matters of homeland security, and the resolution of missing person investigations, through the analysis of found human remains in routine case-work, or in mass disasters. In addition, this comprehensive chip requires the consumption of only one aliquot of DNA evidence material. The proposed methodology will be a valuable complement to crime-scene STR analysis which represents the industry standard for DNA-based identification through direct matching.

Further developments are underway. The current Version 1 of the Identitas Forensic Chip already contains SNPs associated with freckles, moles, curly hair, skin-color, earlobe-shape and body height. However, the phenotype prediction values of the currently-known markers for these EVCs are not high enough to be practically useful: more predictive DNA markers will need to be identified. Subsequent versions of the Identitas Forensic Chip will include additional markers for these and other appearance traits, as they are identified.

Acknowledgements

The VisiGen Consortium wishes to acknowledge the contributions of their research institutions, study investigators, field staff and the study participants, in creating the scientific resources accessed. We also would like to thank the scientists at Illumina for help with assay scoring, assay selection, and genotyping. TwinsUK is supported by the Wellcome Trust and the NHS NIHR Biomedical resource grant to Guys and St Thomas' foundation hospitals and KCL. Identitas Inc. sponsors the VisiGen Consortium via a research grant to a number of the respective academic institutions.

Conflict of Interest

ATB, BK and LR had or have a financial relationship with Identitas Inc. TDS and MK have consulted for Identitas Inc. and are on the SAB but without financial or other direct benefits. All other authors declare that they have no conflict of interest.

Disclaimer

SNP genotyping was supported in part by the FBI Laboratory Division. Names of commercial manufacturers are provided for identification only and inclusion does not imply endorsement of the manufacturer or its products or services by the FBI. The views expressed are those of the authors and do not necessarily reflect the official policy or position of the FBI or the US Government. This manuscript was filed under the number 12-18 at the Counterterrorism and Forensic Science Research Unit, Federal Bureau of Investigation Laboratory Division.

Figure 1. Quantitative assessment of bio-geographic ancestry from an individual whose father was of European origin and whose mother was of Chinese origin

References

1. Spinney L (2008) Eyewitness identification: line-ups on trial. Nature 453 (7194):442-444

2. Wells GL, Malpass RS, Lindsay RC, Fisher RP, Turtle JW, Fulero SM (2000) From the lab to the police station. A successful application of eyewitness research. Am Psychol 55 (6):581-598

3. Spector TD, Williams FM (2006) The UK Adult Twin Registry (TwinsUK). Twin Res Hum Genet 9 (6):899-906

4. Zhu G, Montgomery GW, James MR, Trent JM, Hayward NK, Martin NG, Duffy DL (2007) A genome-wide scan for naevus count: linkage to CDKN2A and to other chromosome regions. Eur J Hum Genet 15 (1):94-102

5. Gu S, Pakstis AJ, Li H, Speed WC, Kidd JR, Kidd KK (2007) Significant variation in haplotype block structure but conservation in tagSNP patterns among global populations. Eur J Hum Genet 15 (3):302-312

6. van Oven M, Kayser M (2009) Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat 30 (2):E386-394

7. Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, Hammer MF (2008) New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Res 18 (5):830-838

8. Chiaroni J, Underhill PA, Cavalli-Sforza LL (2009) Y chromosome diversity, human expansion, drift, and cultural evolution. Proc Natl Acad Sci U S A 106 (48):20174-20179

9. International HapMap Consortium (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467 (7311):52-58

10. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155 (2):945-959

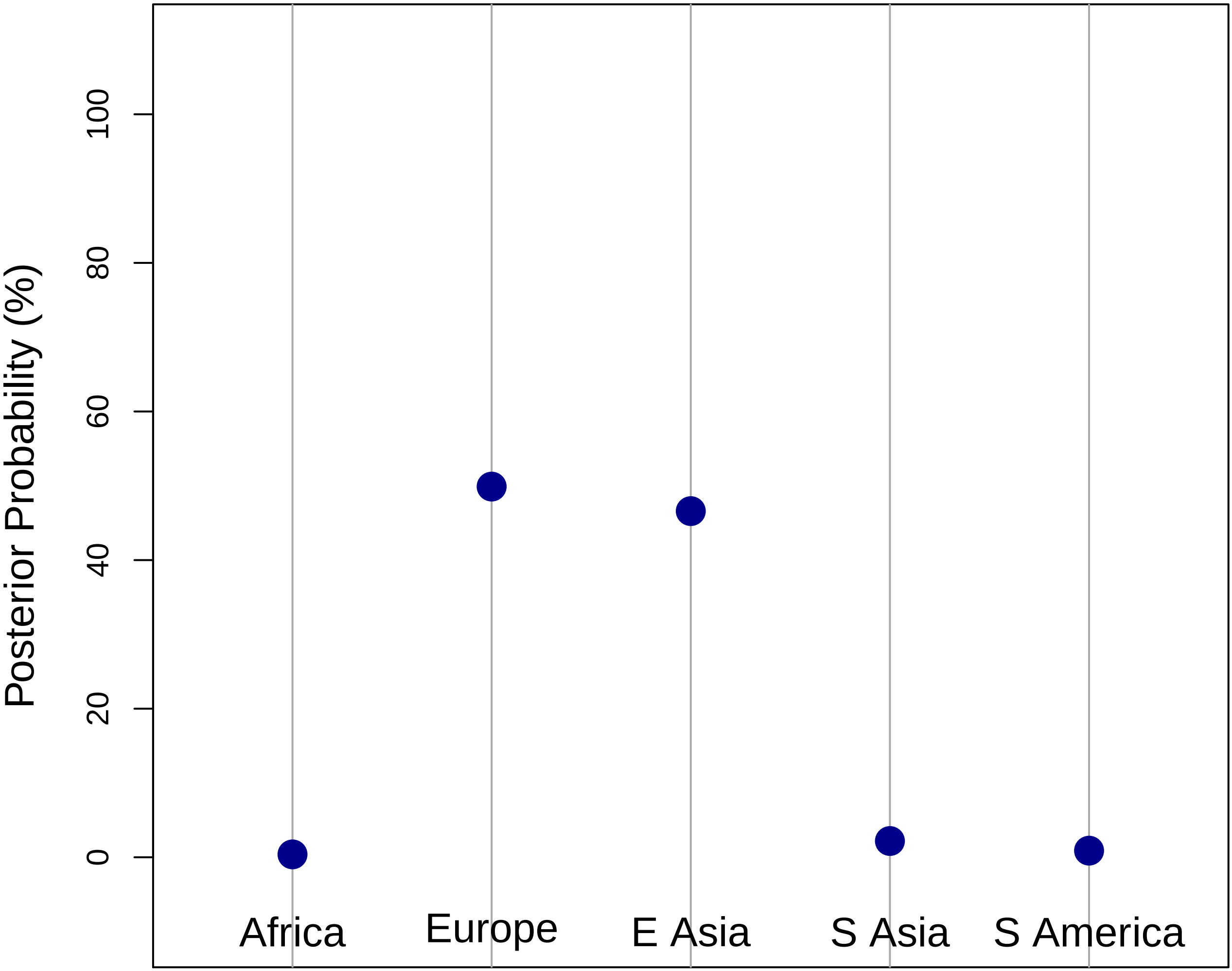
11. Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. Mol Ecol Resour 9 (5):1322-1332

12. Walsh S, Liu F, Wollstein A, Kovatsi L, Ralf A, Kosiniak-Kamysz A, Branicki W, Kayser M (2012) The HIrisPlex System for simultaneous prediction of hair and eye colour from DNA. Forensic Science International: Genetics: http://dx.doi.org/10.1016/j.fsigen.2012.1007.1005

13. Walsh S, Liu F, Ballantyne KN, van Oven M, Lao O, Kayser M (2011) IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. Forensic Sci Int Genet 5 (3):170-180

14. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81 (3):559-575

15. Walsh S, Wollstein A, Liu F, Chakravarthy U, Rahu M, Seland JH, Soubrane G, Tomazzoli L, Topouzis F, Vingerling JR, Vioque J, Fletcher AE, Ballantyne KN, Kayser M (2012) DNA-based eye colour prediction across Europe with the IrisPlex system. Forensic Sci Int Genet 6 (3):330-340


Posterior Probability (%)
0
20
40
60
80
100
Africa
Europe
E Asia
S Asia
S America