"Biological identikit": development of a SNPs-Panel for the analysis of Forensic DNA Phenotyping and Ancestry

Giulia Sguazzi[a,b]*, Debora Varrone[c], Chiara Cirioni[c], Valentina Andrioletti[c], Pasquale Linarello[c], Luca Salvaderi[c], Flavia Lovisolo[a], Noemi Procopio[d], Fabiano Gentile[e], Anna Cherubini[e], Domenico Colloca[e], Alberto Marino[e], Sarah Gino[a]

[a] Department of Health Science, Università del Piemonte Orientale, 28100 Novara, Italy

[b] CRIMEDIM – Center for Research and Training in Disaster Medicine, Humanitarian Aid and Global Health, Università del Piemonte Orientale – 28100 Novara, Italy

[c] Eurofins Genoma – 20161 Milano, Italy

[d] School of Natural Sciences, University of Central Lancashire - Preston, PR1 2HE, UK

[e] Reparto Carabinieri Investigazioni Scientifiche, Parma, Italy

*Corresponding Author: giulia.sguazzi@uniupo.it +39 3393514397

Abstract

Personal identification in mass disasters and in crimes is essential for humanitarian, ethical and legal reasons. In these contexts, when individuals cannot be identified by standard forensic DNA analysis, the Forensic DNA Phenotyping and the analysis of the biogeographical ancestry could help. The aim of this study was to evaluate the potential of a new panel of 891 SNPs in predicting phenotypic traits and biogeographical origin to create a “biological identikit”. In addition to fresh biological material, old evidence found at the crime scene or extracted and long-stored DNA were tested with 41 SNPs for phenotyping and 850 SNPs for ancestry. All the SNPs were successfully incorporated into a single two-step multiplex PCR reaction using the IonAmpliSeq ™ Library Plus and applied for massive parallel sequencing with the Ion S5 platform using up to 0.05 ng / µl of DNA. The analysis of the results was carried out with an in-house predictive algorithm and consulting 20 population databases. By comparing the results obtained with identikit or video-photographic surveys, it was possible to predict phenotype and ancestry with an accuracy greater than 90%. While these new markers cannot identify a specific individual, they can be a valuable investigative tool.

Key words

DNA Phenotyping, Ancestry, SNP, Cold Case, DVI

Introduction

With the advancement of technology, Forensic Genetics laboratories have been more frequently involved in the reopening of unsolved crimes and in the disaster victim identification (DVI), often having to deal with interpretative problems. However, one of the major limitations is the need for a comparative approach that typically prevents to identify persons whose STRs profile is not already known to the investigators [1,2]. Consequently, the “biological witnesses” belonging to unknown suspects remain unused considering also that DNA-based mass screening is not allowed by legislation everywhere.

The identification of victims of mass disasters and perpetrators of crimes is fundamental for humanitarian, ethical and legal reasons [3]. In these contexts, when individuals cannot be identified by standard forensic DNA analysis, the Forensic DNA Phenotyping (FDP) [4,5] and the analysis of Biogeographic Ancestry (BGA) [1,6] can provide more accurate and reliable testimonies than eyewitness. In a context such as DVI, they could provide the investigator with a high number of information on the victim, thus creating a kind of "biological identikit", useful in the identification process and in the reassembly of the remains [7].

Material and Methods

We selected 30 DNA samples (five reference samples (saliva) and 25 biological evidence (blood and semen)) collected at the crime scenes between 2016 and 2021: the DNA was extracted at the time of the investigations, quantified with Quantifiler™ Trio DNA Quantification Kit and subsequent subjected to genetic typing. In addition, eight recent DNA samples taken from subjects, whose phenotypic characteristics and ancestral origin were known, were also analysed. The quantification was repeated, for all 38 samples, using the Qubit Fluorometric Quantification.

Afterwards only 34 DNA extracts were subjected to Massive Parallel Sequencing (MPS) using a panel of 891 SNPs: 41 SNPs for FDP and 850 SNPs for BGA. All the SNPs were successfully incorporated into a single two-step multiplex PCR reaction using the IonAmpliSeq™ Library Plus and applied for MPS with the Ion S5 platform using up to 0.05 ng/µl of DNA. The analysis of the results was carried out non only through the HirisPlex-S Webtool (https://hirisplex.erasmusmc.nl/) to generate individual prediction probabilities for three eye colours, four hair colours and five skin colour categories, but also with an in-house predictive algorithm and consulting 20 databases containing population frequencies.

Results and Discussion

Discrepancies between the two quantifications (current and at the time of the investigation) were found for the 30 archival samples: the DNA concentration appears underestimated in comparison with the initial quantification (**Table 1**).

Sample	Biological evidences	Quantifiler trio (ng/ul)			Qubit (ng/ul)	Dna loaded (ul)	Library (ng/ml)
		Human	Male	Degradate			
A1	*Blood*	*2.98*	*4.27*	*2.12*	*0.698*	*12*	*89*
B1	Semen	27.3	36.13	0.81	11.80	12 (2ng/ul)	7440
C1	Blood	2.42	na	1.06	0.169	12	4440
D1	Blood	2.37	1.35	1.71	1.13	12	5040
E1	Blood	1.48	1.005	0.87	0.511	12	4880
F1	Blood	2.36	1.83	1.64	0.89	12	5360
G1	Blood	2.04	2.3	0.48	0.183	12	5240
H1	Blood	3.27	5.24	0.51	0.303	12	5520
A2	*Blood*	*2.79*	*2.7*	*0.57*	*0.262*	*12*	*72*
B2	*Blood*	*2.43*	*2.28*	*1.52*	*1.08*	*12*	*250*
C2	Blood	3.39	2.89	1.5	1.04	12	5080
D2	Blood	1.13	0.76	0.49	0.263	12	3952
E2	Blood	2.52	2.15	1.97	0.708	12	4720
F2	Blood	2.78	2.3	1.004	1.14	12	5280
G2	Blood	1.54	1.06	0.86	0.157	12	3996
H2	Blood	2.15	2.1	0.85	0.562	12	5080
A3	*Blood*	*0.10*	*0.094*	*0.86*	*too low*		
B3	Blood	1.18	0.94	0.7	0.572	12	5400
C3	Blood	1.02	1.05	0.96	0.558	12	5240
D3	Blood	1.68	1.73	0.85	0.965	12	4760
A9	Blood	1.33	1.78	0.9	1.05	12	5280
B9	Blood	2.29	2.75	0.92	1.63	12	4840
C9	Blood	4.30	3.8	1.14	2.38	12	5240
D9	Blood	3.62	3.79	0.7	2.04	12	4720
E9	Saliva	17.88	0	1.32	6.60	12	7120
F9	Saliva	11.76	0	1.17	5.34	12	6960
G9	Saliva	9.35	12.25	1.04	1.75	12	5120
H9	Saliva	10.52	12.68	0.86	8.91	12 (2ng/ul)	5200
A10	Saliva	4.80	na	0.83	3.12	12	5800

B10	Saliva	14.47	0	0.87	15.2	12 (2ng/ul)	6080
1	Saliva	-	-	-	15.7	12 (2ng/ul)	1612
2	Saliva	-	-	-	n.a	12 (2ng/ul)	2176
3	Saliva	-	-	-	13.1	12 (2ng/ul)	1806
4	Saliva	-	-	-	9.11	12 (2ng/ul)	2464
5	Saliva	-	-	-	26.1	12 (2ng/ul)	1436
6	Saliva	-	-	-	n.a	12 (2ng/ul)	1352
7	Saliva	-	-	-	n.a	12 (2ng/ul)	820
8	Saliva	-	-	-	n.a	12 (2ng/ul)	1244

Table 1. Quantification results. Samples excluded from analysis are reported in cursive.

The MPS results showed a good performance of the designed panel. All SNPs have been uniformly amplified and sequenced in the different types of samples, without differences between reference and degraded samples and regardless of the amplicon size and the degradation rate. Only four out of the 30 archival samples (A1,A2,A3,B2) have been excluded from the library because the low DNA concentration or the reduced library (**Table 1**). This confirm that the degradation index does not influence the correct genotyping, whilst the critical parameter that affect the result seems to be the quantity of input DNA.

A sensitivity test was conducted to determine the minimum input DNA needed to obtain a complete 891 SNPs profile. Analysing the area under the operating curves (AUC) as overall measure of prediction accuracy by using a reduced number of sample for which ancestry and phenotype were known, complete and reliable predictions were obtained also with DNA concentrations as low as 1.02 ng/ul. Particularly emblematic is the case of E2 and G2 samples (blood evidence collected from the same crime scene that share the same STR profile): an identical p-values was obtained when predicting both phenotypic characteristics and biogeographic origins.

Concordant results between predicted and expected phenotypes were also obtained for reference samples. For these samples there were no problems in the prediction of hair and skin colours, however it was more difficult to predict the colour of the eyes, especially for the intermediate tones. Regarding BGA, when applying our Panel on a subcontinental level to European populations in the 1000 genomes dataset [8], we identified at a novel 850 ancestral informative markers set (AIMs), that numerically exceeded all other panels available and provided accurate predictions. However, passing by a sub-continental level to sub-regional one, the misclassification error did not drop below 50%. We therefore consider that going into too much detail could add irrelevant information for the forensic geneticist, interested in distinguishing geographical origins characterized by a different physiognomy. Another complication in the correct prediction of BGA lies in the real definition of the ancestral origin of the subject. Typically, this information is provided by the test subject, and as such, can be ambiguous, wrong, or not entirely known.

Conclusion

This study showed the possibility to apply this Panel to crime scene evidence and in mass disaster for personal identification purposes. The results highlighted that complete and accurate phenotypic prediction were possible even from 100pg of degraded DNA. Despite, historically the DNA profiling has involved comparison with ante mortem samples or relatives, now it can direct investigators towards putative victims or relatives for comparison, through the determination of externally visible characteristics and ancestry. Obviously further validation studies with additional samples will be needed for better assessments on its effectiveness and usefulness in forensic caseworks.

Acknowledgements

This article is the result of a study conducted in the framework of the International PhD in Global Health, Humanitarian Aid, and Disaster Medicine organized by University of Eastern Piedmont (UPO).

Declaration of Competing Interest

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of this paper.

References

[1] P.M. Schneider, B. Prainsack, M. Kayser, The use of forensic DNA phenotyping in predicting appearance and biogeographic ancestry, Dtsch. Arztebl. Int. 116 (2019) 873–880. https://doi.org/10.3238/arztebl.2019.0873.

[2] M. Kayser, Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes, Forensic Sci. Int. Genet. 18 (2015) 33–48. https://doi.org/10.1016/j.fsigen.2015.02.003.

[3] L. Caenazzo, S. Gino, La medicina legale nella protezione dei diritti umani, (2021). http://www.piccin.it/it/medicina-legale/2536-la-medicina-legale-nella-protezione-dei-diritti-umani-9788829930982.html (accessed February 7, 2022).

[4] M. Kayser, Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes, Forensic Sci. Int. Genet. 18 (2015) 33–48. https://doi.org/10.1016/j.fsigen.2015.02.003.

[5] M. Kayser, P.M. Schneider, DNA-based prediction of human externally visible characteristics in forensics: Motivations, scientific challenges, and ethical considerations, Forensic Sci. Int. Genet. 3 (2009) 154–161. https://doi.org/https://doi.org/10.1016/j.fsigen.2009.01.012.

[6] C. Xavier, M. de la Puente, A. Mosquera-Miguel, A. Freire-Aradas, V. Kalamara, A. Vidaki, T. E. Gross, A. Revoir, E. Pośpiech, E. Kartasińska, M. Spólnicka, W. Branicki, C. E. Ames, P. M. Schneider, C. Hohoff, M. Kayser, C. Phillips, W. Parson, Development and validation of the VISAGE AmpliSeq basic tool to predict appearance and ancestry from DNA, Forensic Sci. Int. Genet. 48 (2020) 102336. https://doi.org/10.1016/j.fsigen.2020.102336.

[7] L. Chaitanya, I.Z. Pajnič, S. Walsh, J. Balažic, T. Zupanc, M. Kayser, Bringing colour back after 70 years: Predicting eye and hair colour from skeletal remains of World War II victims using the HIrisPlex system, Forensic Sci. Int. Genet. 26 (2017) 48–57. https://doi.org/10.1016/j.fsigen.2016.10.004.

[8] S. Fairley, E. Lowy-Gallego, E. Perry, P. Flicek, The International Genome Sample Resource (IGSR) collection of open human genomic variation resources, Nucleic Acids Res. 48 (2020) D941–D947. https://doi.org/10.1093/nar/gkz836.