Aspects of Phenotyping and Ancestry – and How Should We Report?

Denise Syndercombe Court
Professor of Forensic Genetics
King's Forensics

VEROGEN

KING'S College LONDON

DEII 2018

Outline – Phenotypics in King's Forensics

Age

- Current capabilities and continuing research

Skin colour

- Applying the HIrisPlex-S snps for skin colour

Biogeographical ancestry

- Use in the UK population
- Using sequence variants

Reporting

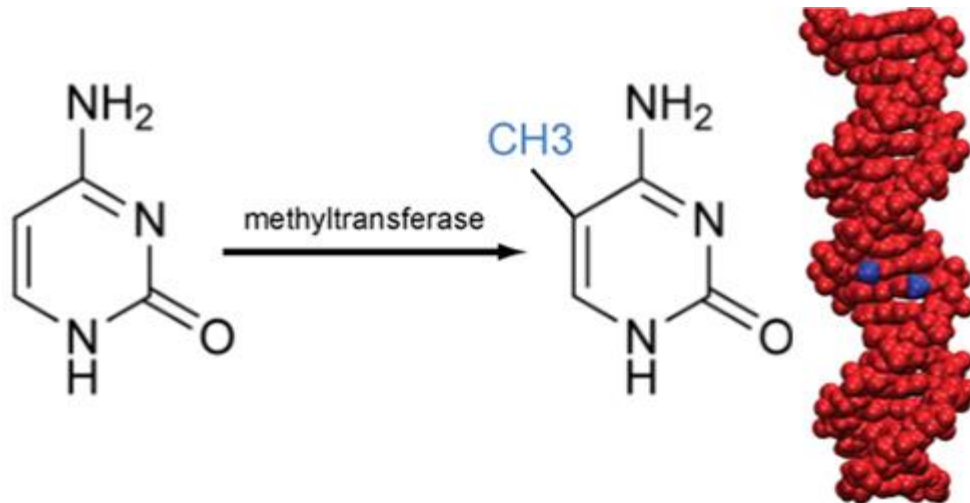- Different approaches

Conclusions

AGE – epigenetic tools for blood stains
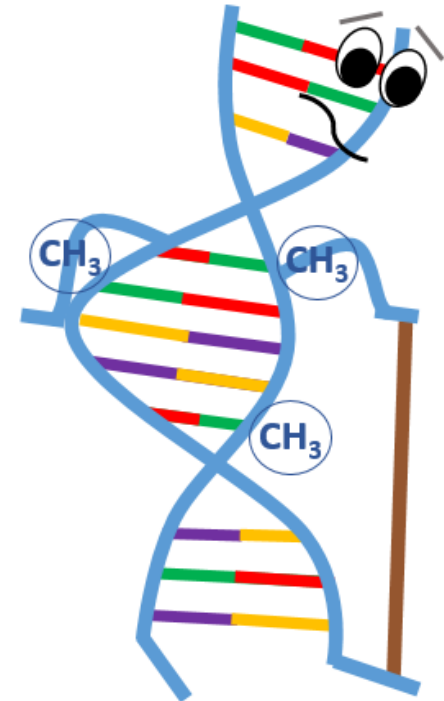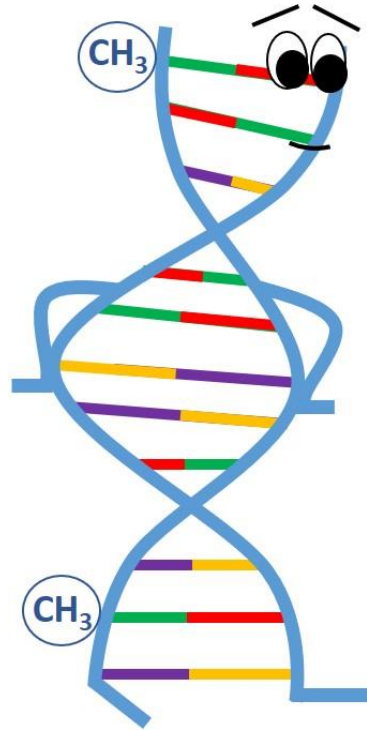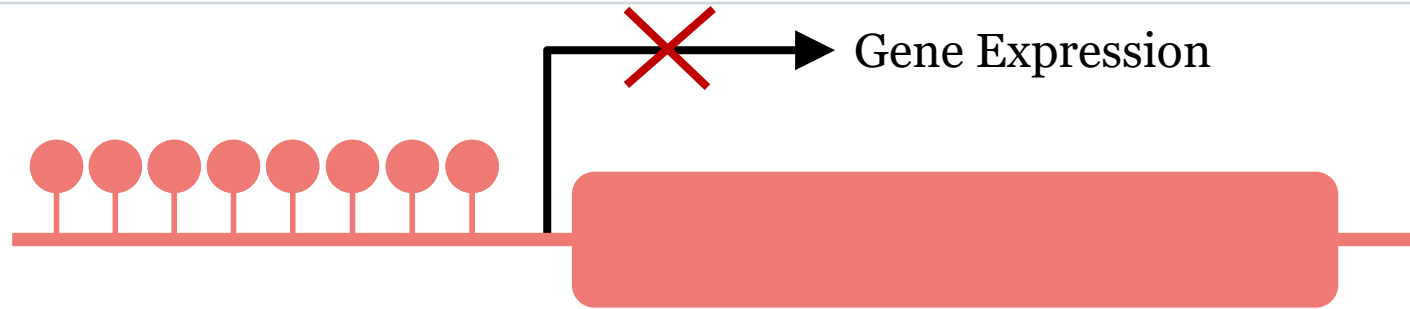
Anastasia Aliferi

Age prediction – different strategies

- Telomere shortening
- Genetic rearrangements
- Epigenetic factors
  - Histone modifications
  - Nucleosomal remodelling
  - Chromatin looping
  - Non coding RNAs
  - DNA methylation

DNA Methylation in Age Prediction

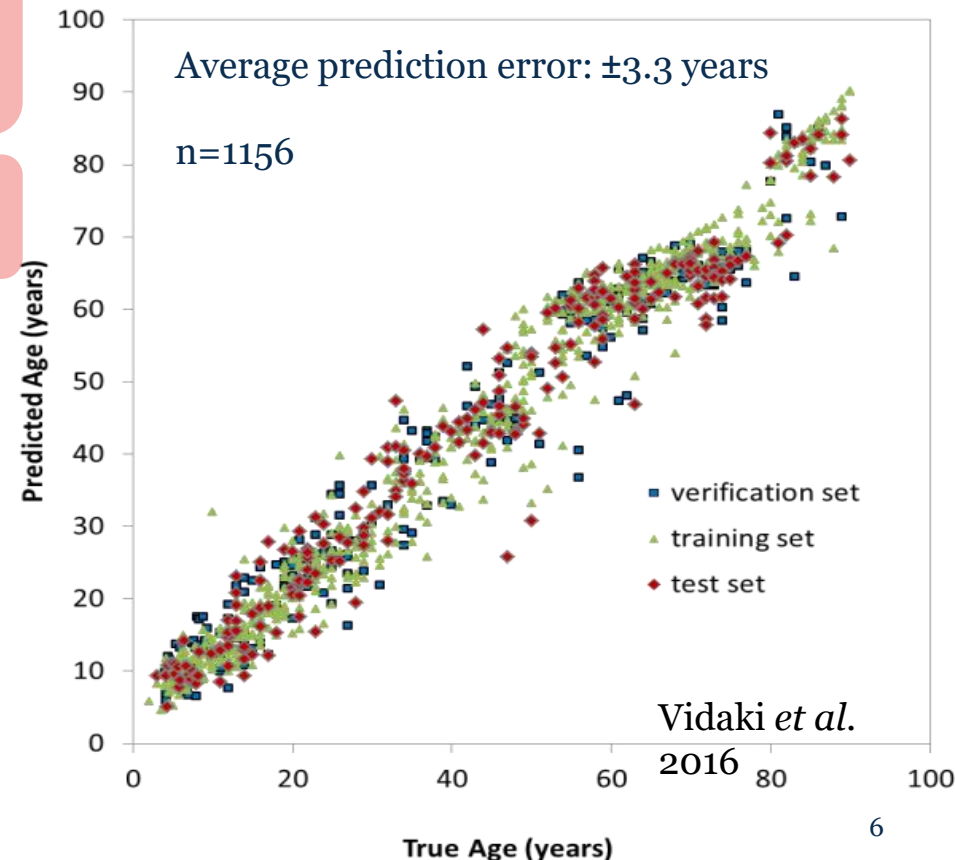# Initial Prediction Model

353 CpGs '**Epigenetic Clock**' - Horvath 2013

12 age-related CpGs

Publicly available **methylation data** deriving from the Illumina 27K and 450K BeadChip arrays

Statistical modelling using artificial neural networks

Average prediction error: ±3.3 years

n=1156

verification set
training set
test set

Predicted Age (years)

True Age (years)
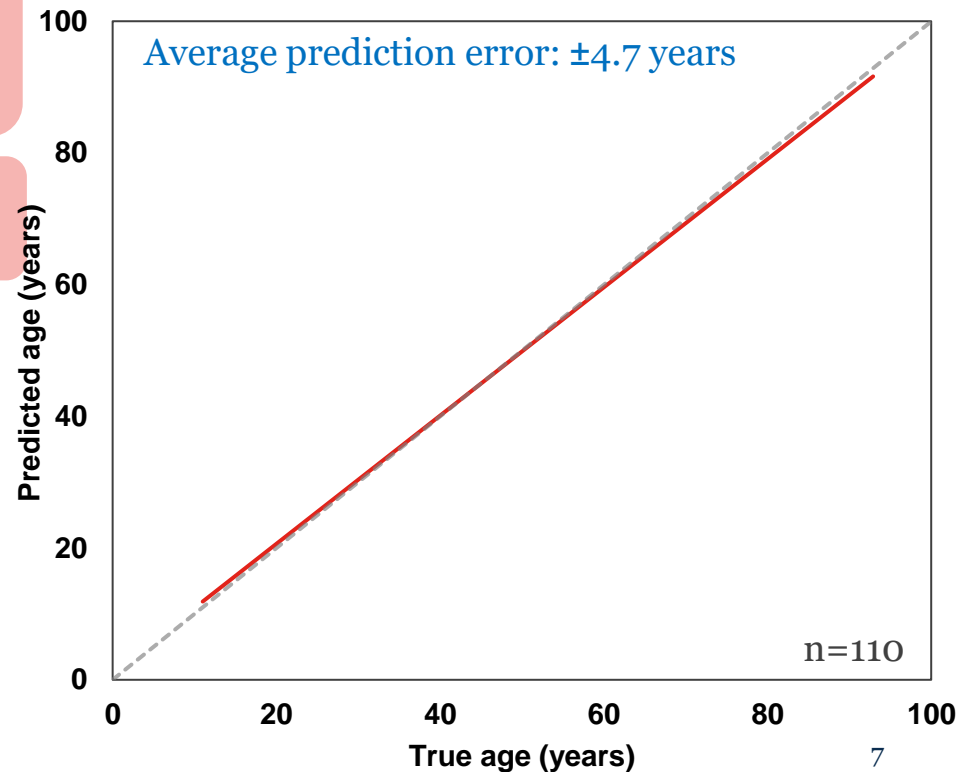
Vidaki *et al.* 2016

Applying the method to real samples

110 whole blood samples

Targeted DNA methylation quantification using the Illumina MiSeq technology (50ng/sample)

67 samples used for **training**, 10 for **validation** and 33 for **blind testing**

Average prediction error: ±4.7 years
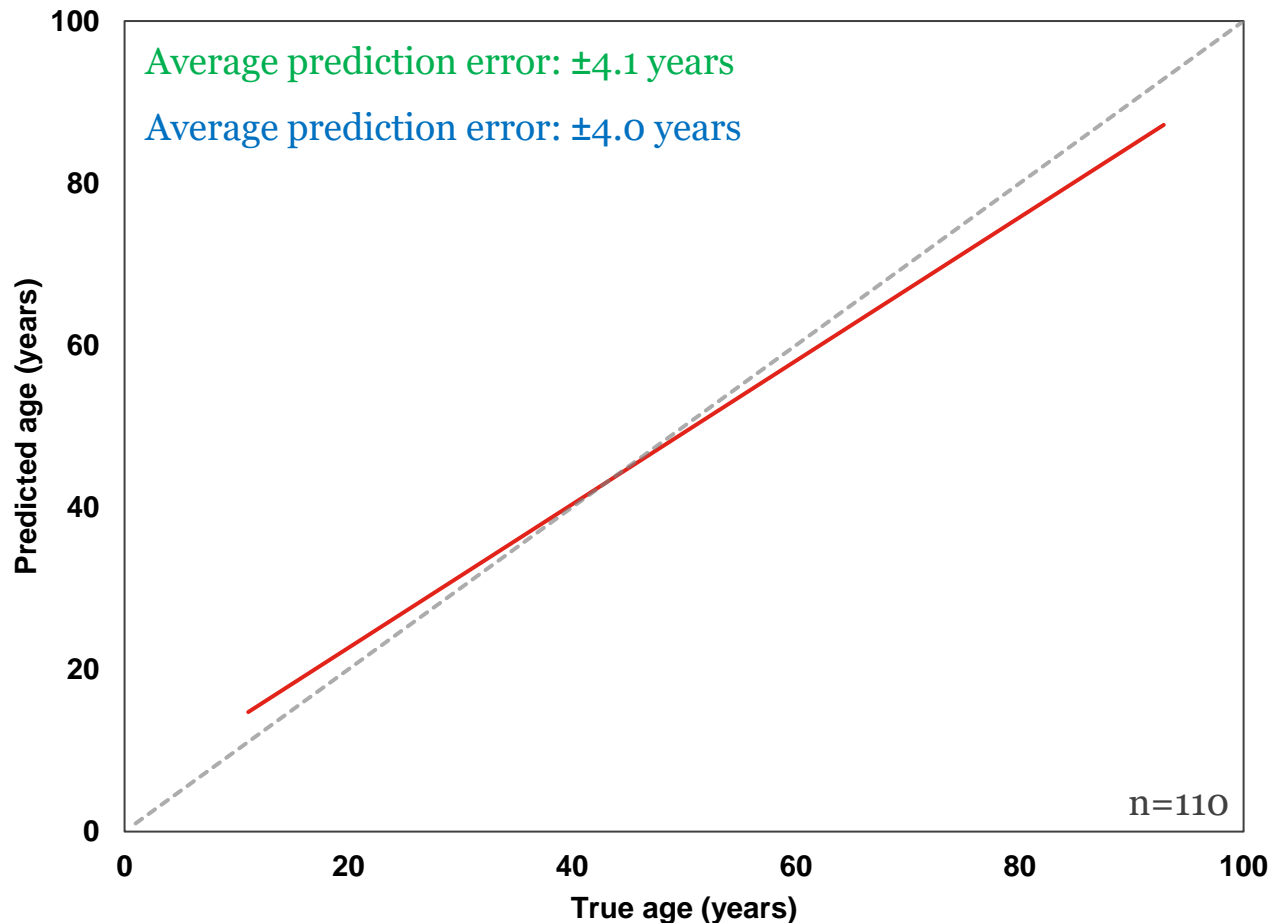

Average prediction error: ±4.7 years

n=110

Statistical modelling

16 additional statistical approaches were tested together with the initial generalised regression model.

A support vector machine model with polynomial function was selected

Average prediction error: ±4.1 years

Average prediction error: ±4.0 years

n=110

Predicted age (years)

True age (years)

Statistical modelling

52% of the samples in the blind test predicting with **an absolute error of less than 4 years** and **85% with an error of less than 7 years**
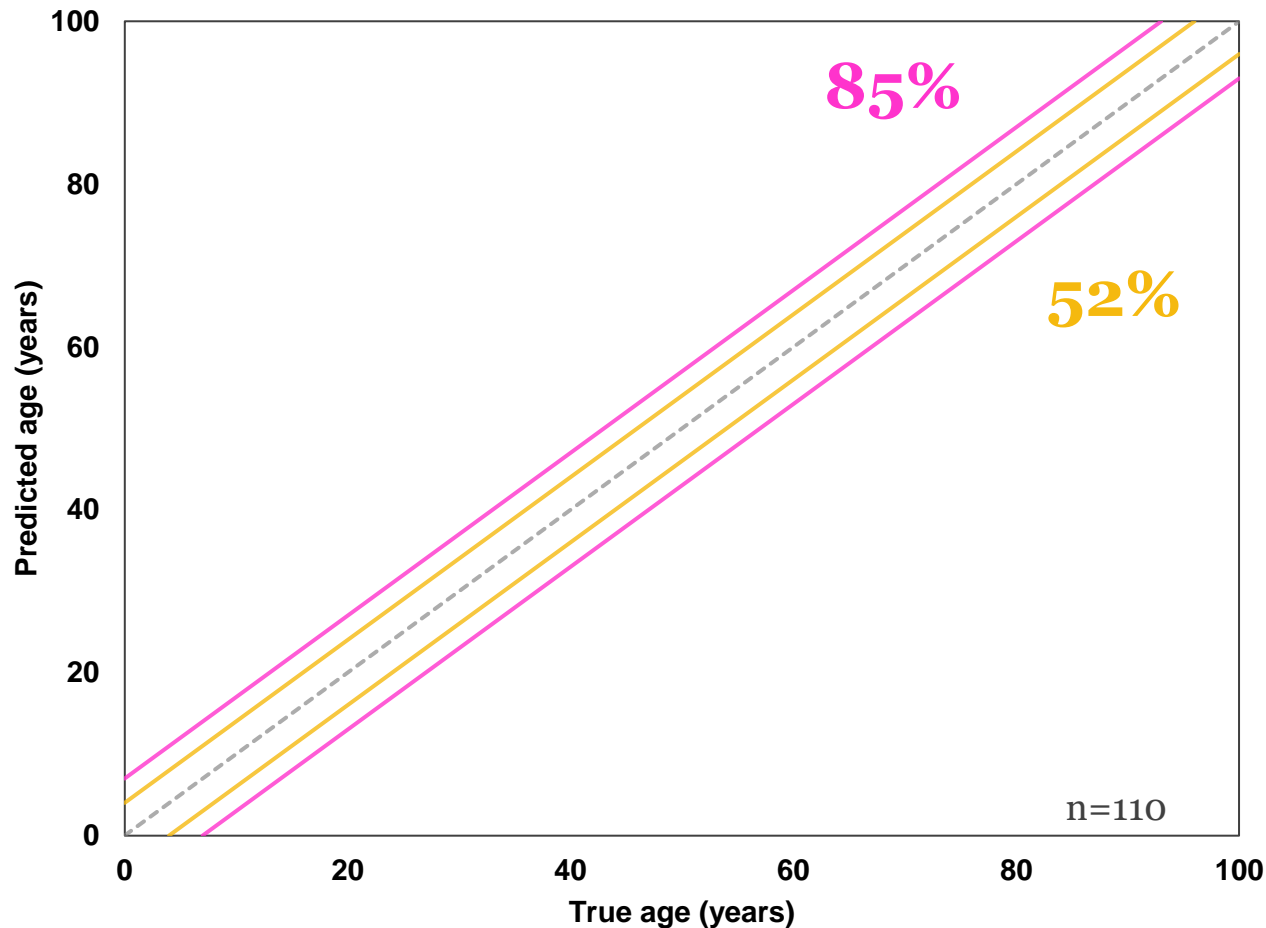
Sensitivity

The **sensitivity** of the method was also assessed using real samples. Below **10ng** the variation increases significantly.
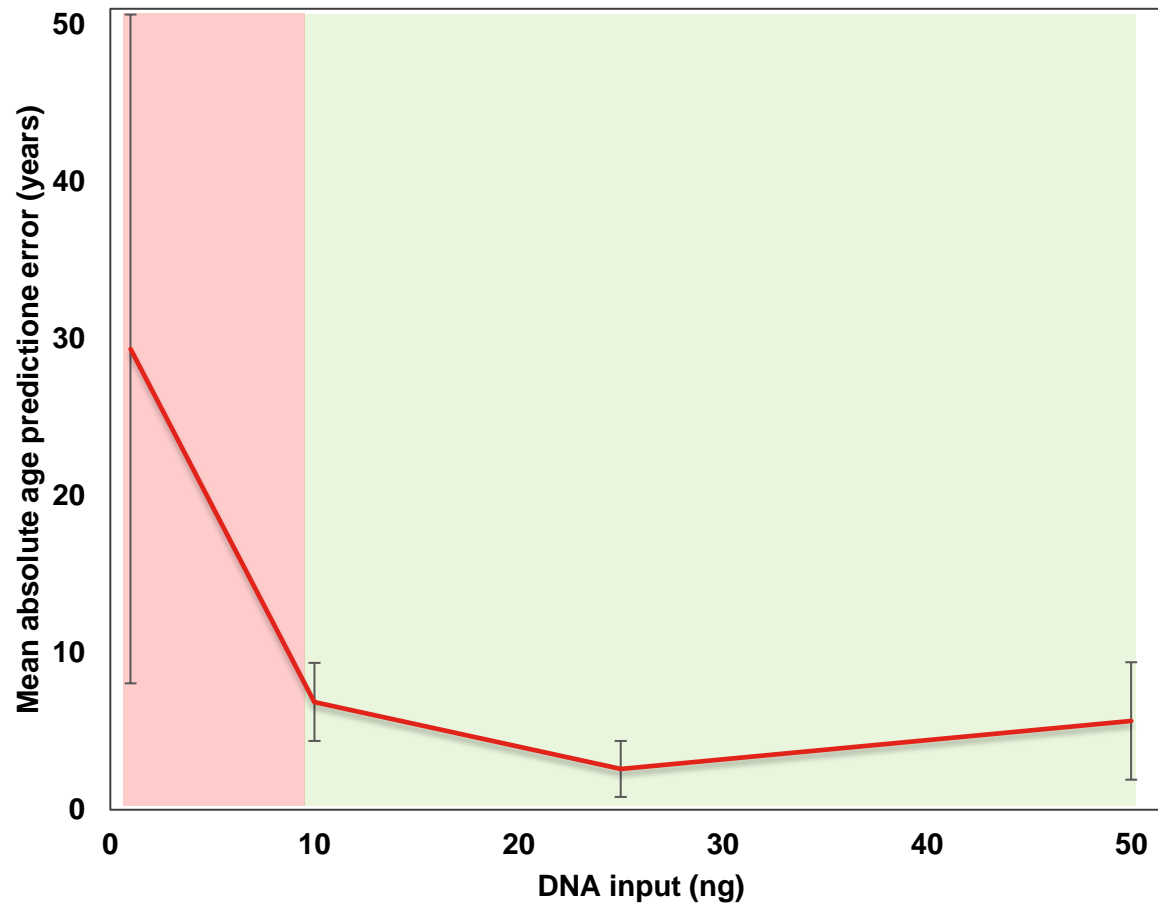
New age-related CpG sites:

from >6000 in 1 or more studies or prediction models

Top 11 CpGs

Effect of sex

Effect of disease

Effect of geographical ancestry

Skin colour – using HIrisPlex-S snps

Ruth Hatton

Skin colour – making use of HIrisPlex-S

Research paper

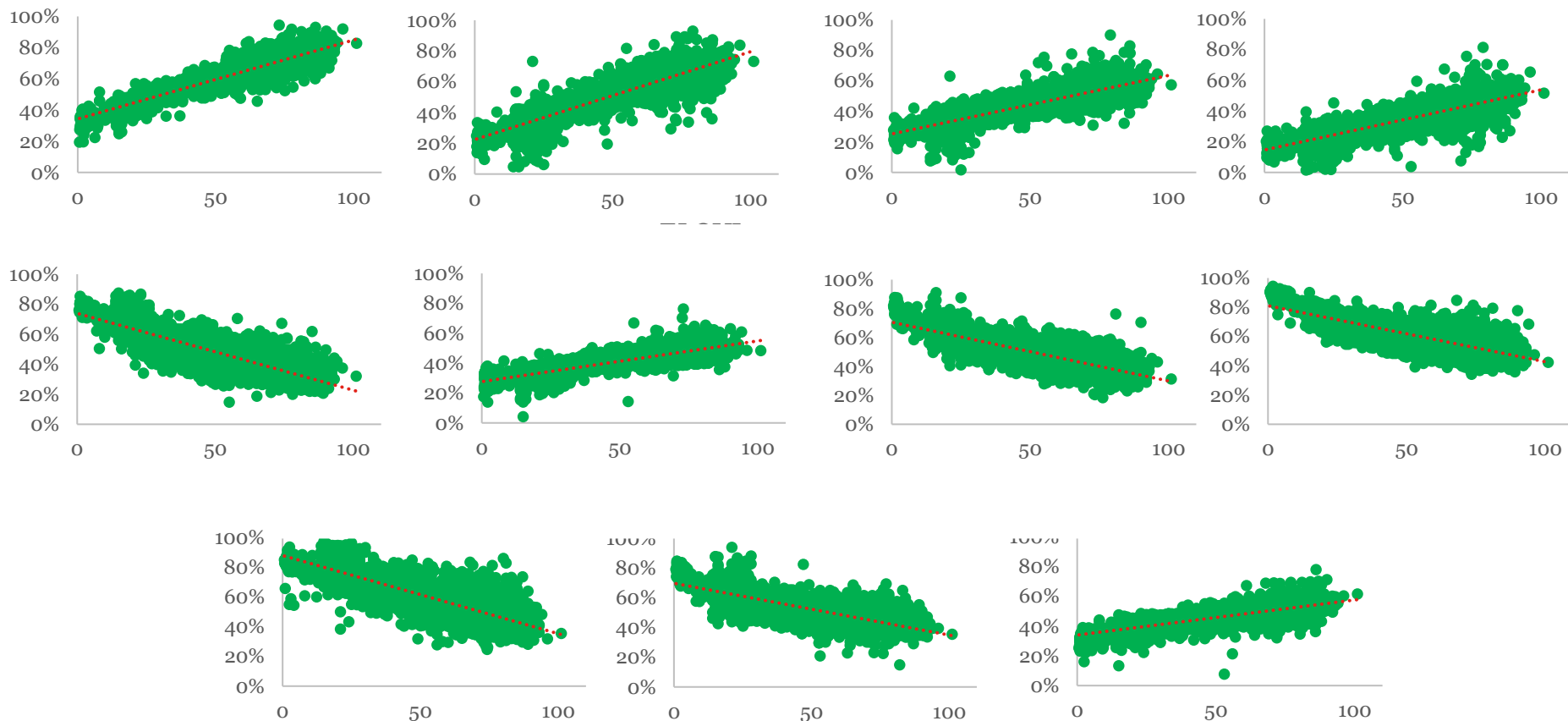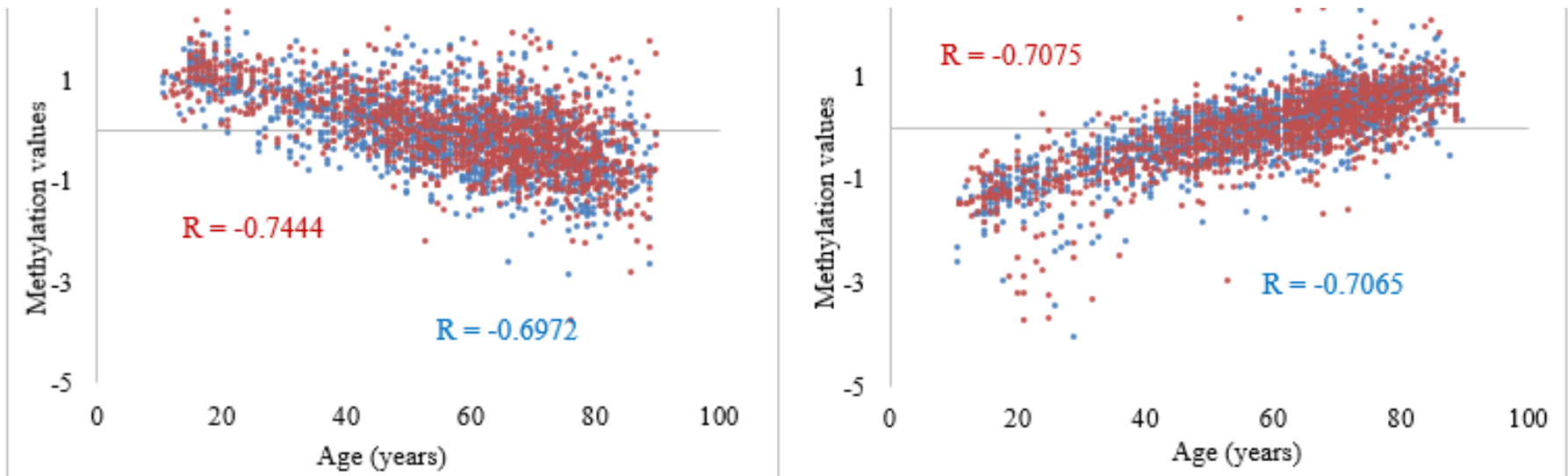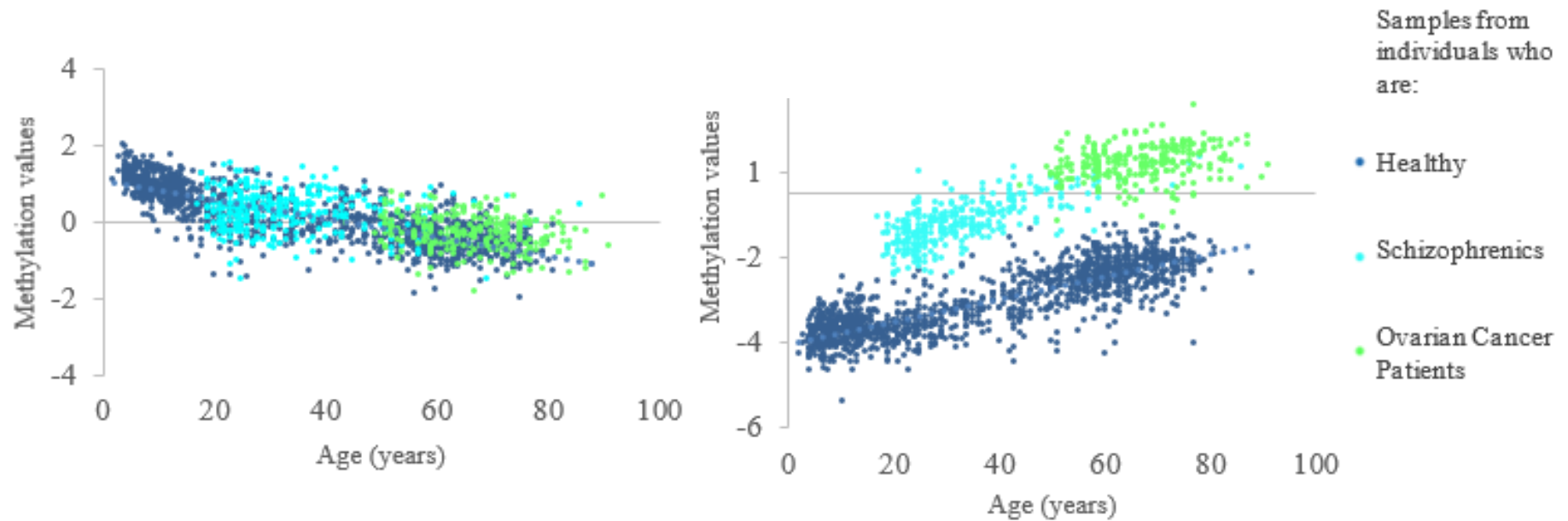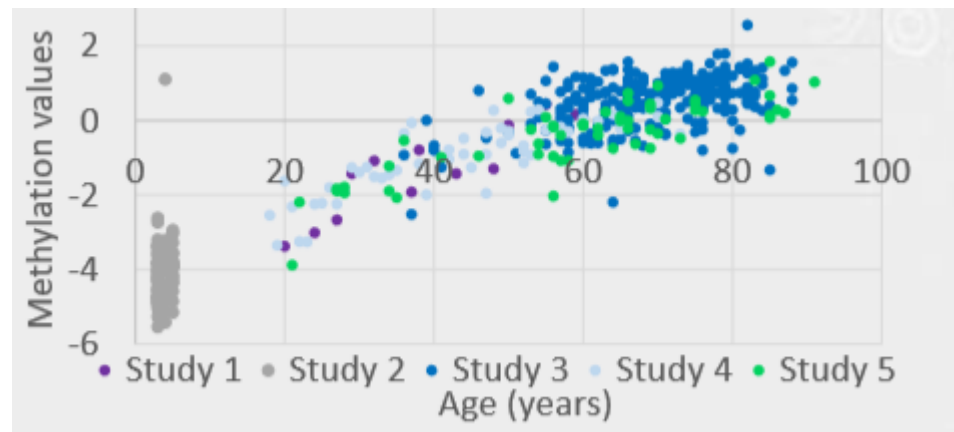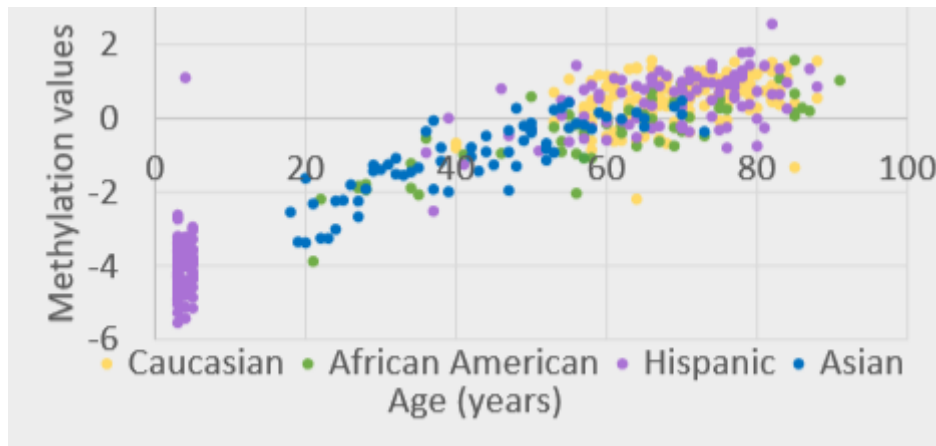# The HIrisPlex-S system for eye, hair and skin colour prediction from DNA: Introduction and forensic developmental validation

Lakshmi Chaitanya[a], Krystal Breslin[b], Sofia Zuñiga[c], Laura Wirken[d], Ewelina Pośpiech[e], Magdalena Kukla-Bartoszek[f], Titia Sijen[c], Peter de Knijff[d], Fan Liu[a,g,h], Wojciech Branicki[e,i], Manfred Kayser[a,*,1], Susan Walsh[b,*,1]

[a] Department of Genetic Identification, Erasmus MC University Medical Centre Rotterdam, Rotterdam, The Netherlands
[b] Department of Biology, Indiana University Purdue University Indianapolis (IUPUI), IN, USA
[c] Division Biological Traces, Netherlands Forensic Institute, The Hague, The Netherlands
[d] Forensic Laboratory for DNA Research, Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands
[e] Malopolska Centre of Biotechnology, Jagiellonian University, Kraków, Poland
[f] Faculty of Biochemistry, Biophysics and Biotechnology of the Jagiellonian University, Kraków, Poland
[g] Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China
[h] University of Chinese Academy of Sciences, Beijing, China
[i] Central Forensic Laboratory of the Police, Warsaw, Poland

HIrisPlex-S Eye, Hair and Skin Colour DNA Phenotyping Webtool
https://hirisplex.erasmusmc.nl/ ▼

Skin study

Biogeographical Origin of Participants

- 15.00% UK
- 7.50% Europe
- 5.00% W. Asia
- 10.00% E. Asia
- 22.50% S. Asia
- 2.50% USA
- 5.00% S. America
- 7.50% NE. Africa
- 10.00% W. Africa
- 7.50% E. Africa
- 7.50% Mixed
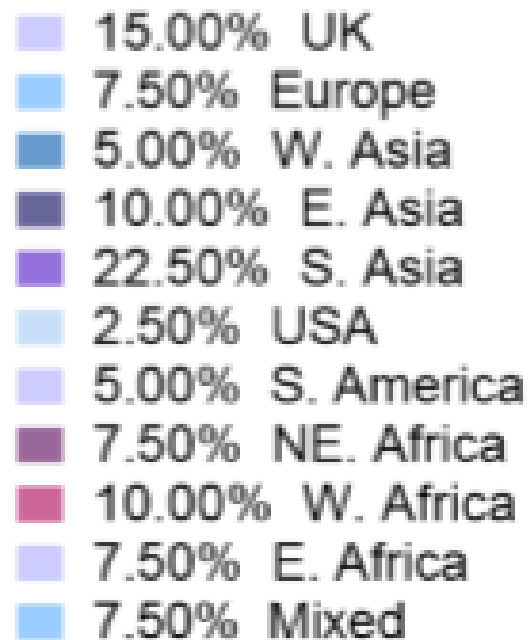
Total = 40 Participants

Skin colour density

L*average values by Biogeographical Origin
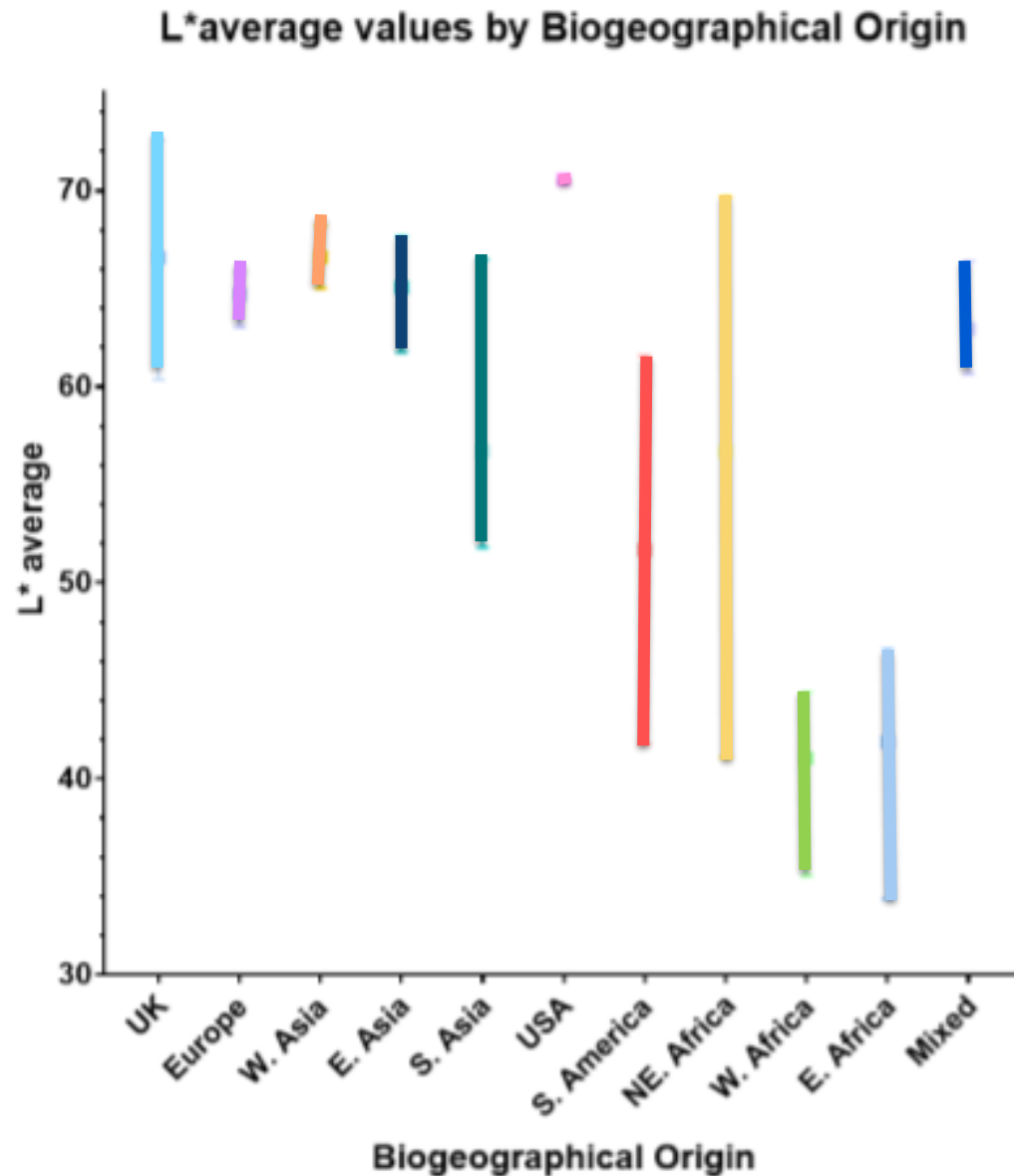
Skin colour scales

	1 Always burns Never tans	2 Usually burns Minimal tan	3 Mild burn Tans uniformly	4 Minimal burn Always tans well	5 Very rarely burns Tans very easily	6 Never burns
Fitzpatrick Scale						
Von Luschan Scale	0 - 6	7 - 13	14 - 20	21 - 27	28 - 34	35 - 36
HIrisPlex-S Scale	Very Pale	Pale	Intermediate		Dark	Dark-Black
Participant Example						

Fitzpatick score v densitometer readings

Using the web tool

A	SKIN	Probability
	Very Pale	0.101
	Pale	0.117
	Intermediate	0.782
	Dark	0
	Dark-Black	0

E	SKIN	Probability
	Very Pale	0
	Pale	0
	Intermediate	0
	Dark	0.001
	Dark-Black	0.999

B	SKIN	Probability
	Very Pale	0
	Pale	0
	Intermediate	0.02
	Dark	0.509
	Dark-Black	0.471

F	SKIN	Probability
	Very Pale	0.102
	Pale	0.25
	Intermediate	0.645
	Dark	0.002
	Dark-Black	0.002

C	SKIN	Probability
	Very Pale	0
	Pale	0
	Intermediate	0
	Dark	0.001
	Dark-Black	0.999

G	SKIN	Probability
	Very Pale	0.003
	Pale	0.011
	Intermediate	0.303
	Dark	0.205
	Dark-Black	0.478

D	SKIN	Probability
	Very Pale	0.005
	Pale	0.364
	Intermediate	0.607
	Dark	0.02
	Dark-Black	0.004

H	SKIN	Probability
	Very Pale	0
	Pale	0
	Intermediate	0.039
	Dark	0.817
	Dark-Black	0.144

Biogeographic ancestry

David Ballard and Laurence Devesse

2011 census – London

Percentage of population

■ White ■ Black ■ South Asian ■ Other

59%

13%

15%

13%
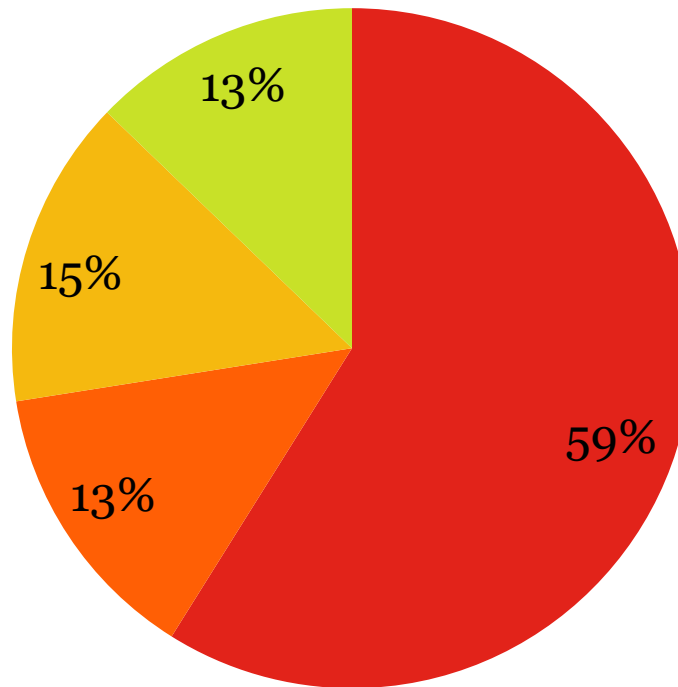
King's Ancestry SNP Panel
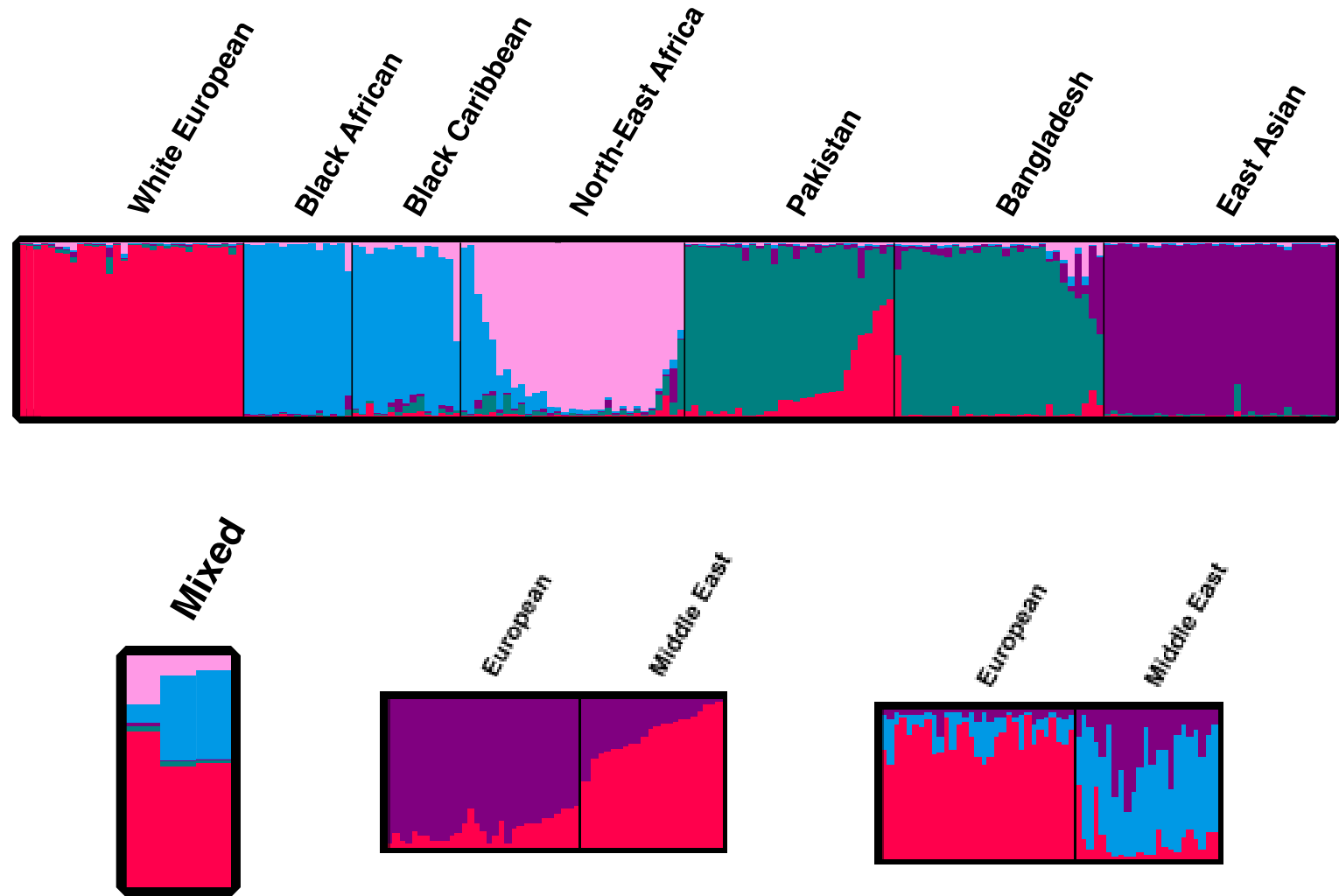
Using STRs

Using STRs for ancestry estimation

- Rosenberg (2016) showed that the standard 13 CODIS markers have ancestry information comparable to a random marker set
- Much larger marker set required to adequately differentiate continental populations
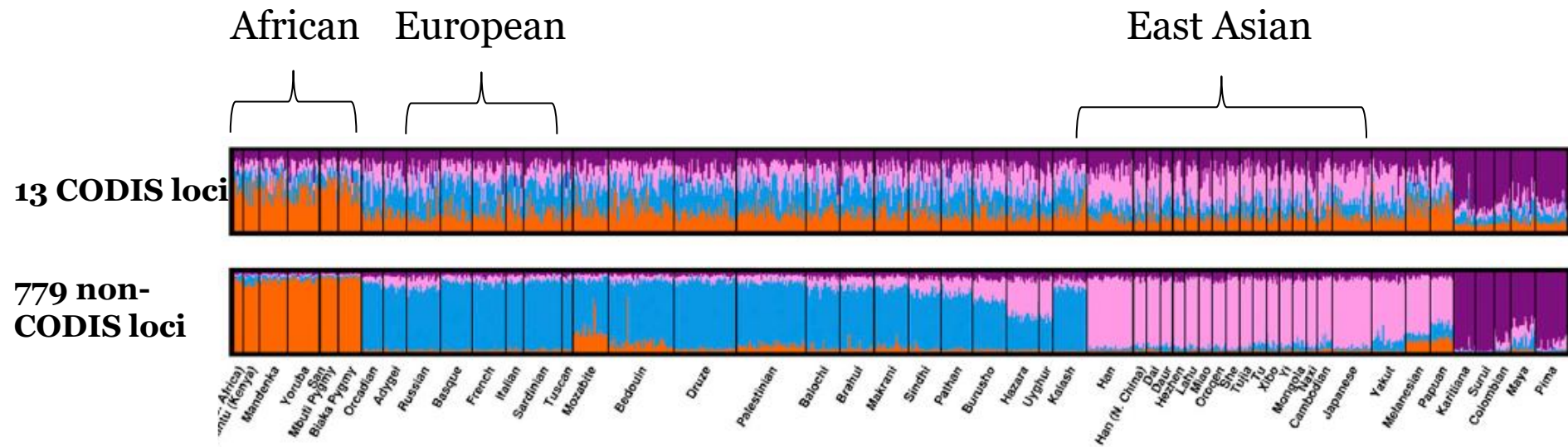
Sequence variation
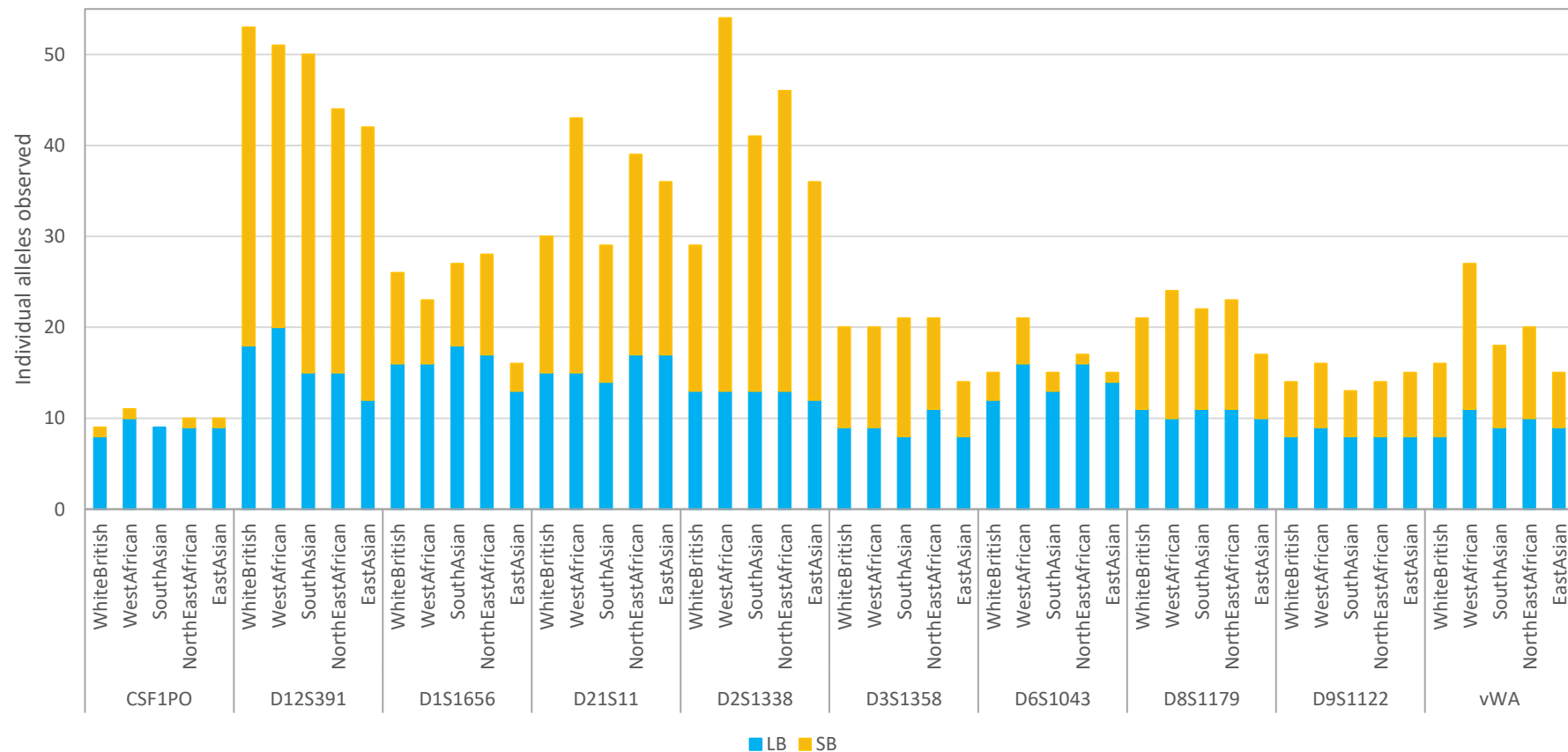
Flanking region variation

- Variation has also been observed in the flanking region of STRs
- Data from two population groups: White European and South Asian
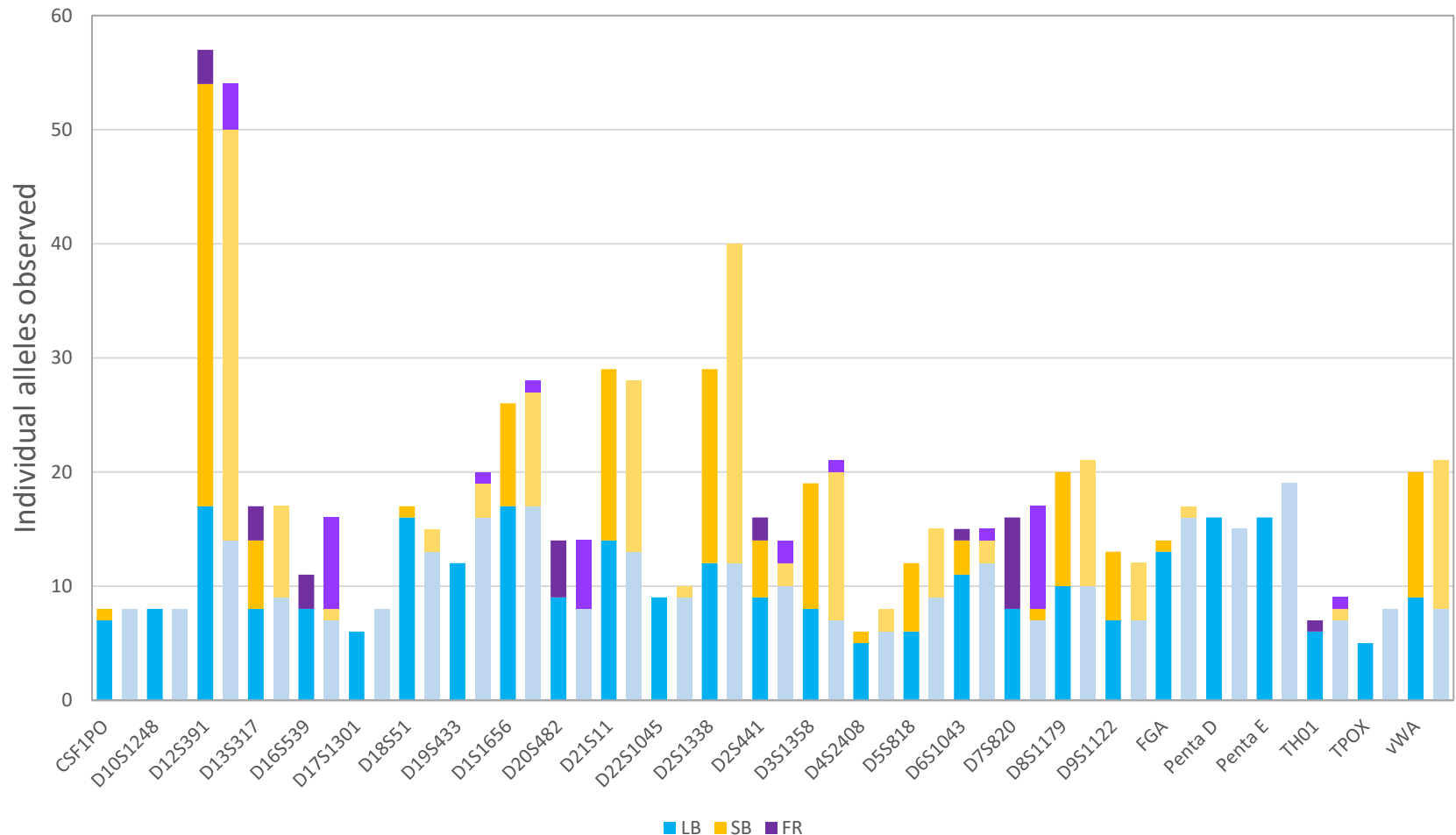
Population specific variation
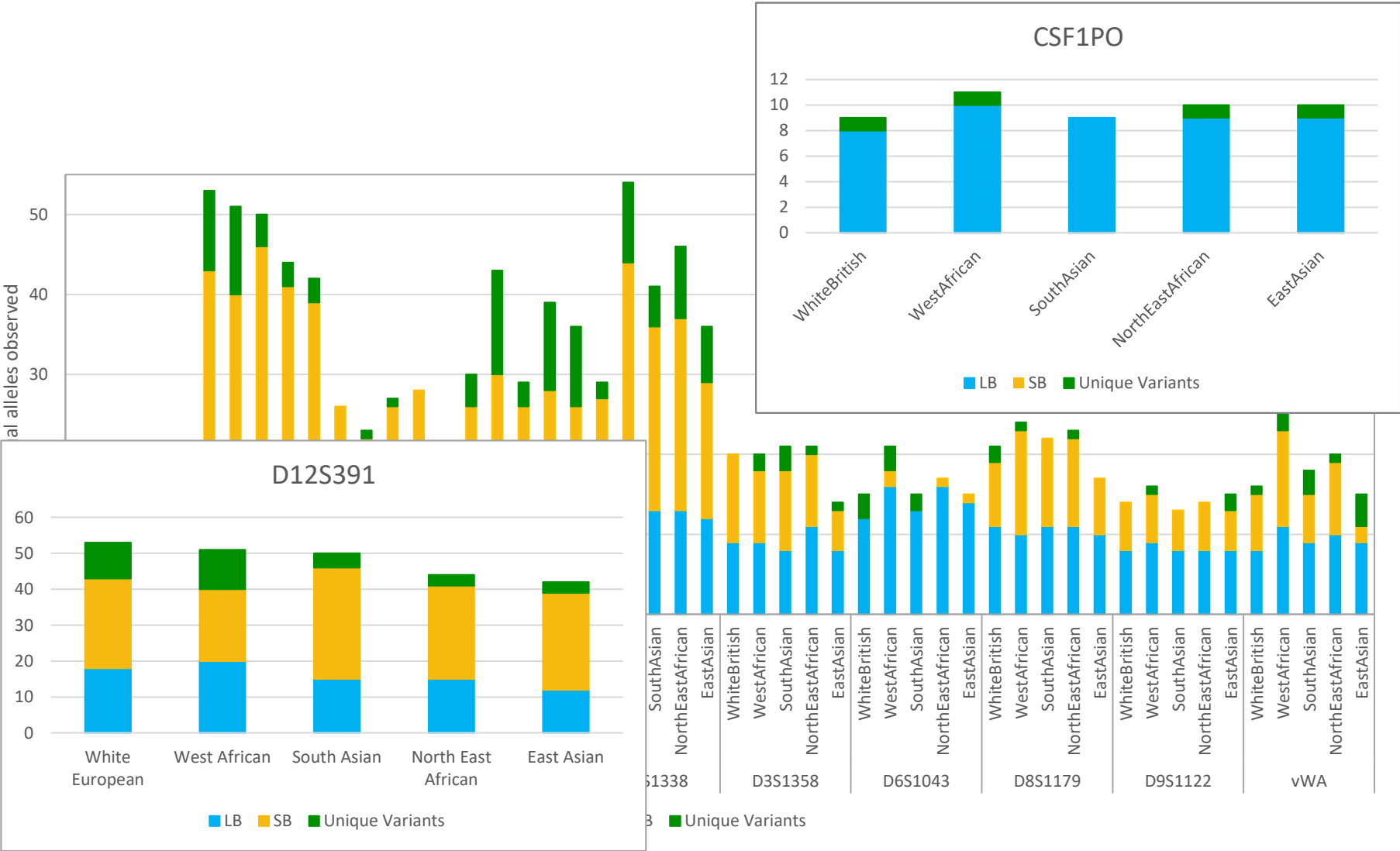
Bio-geographical ancestry inference

- Can we use sequence variation to infer ancestry?

Length-based alleles (CE) K 3

White European   West African   South Asian   North East African   East Asian

Sequence-based alleles (MPS) K 4
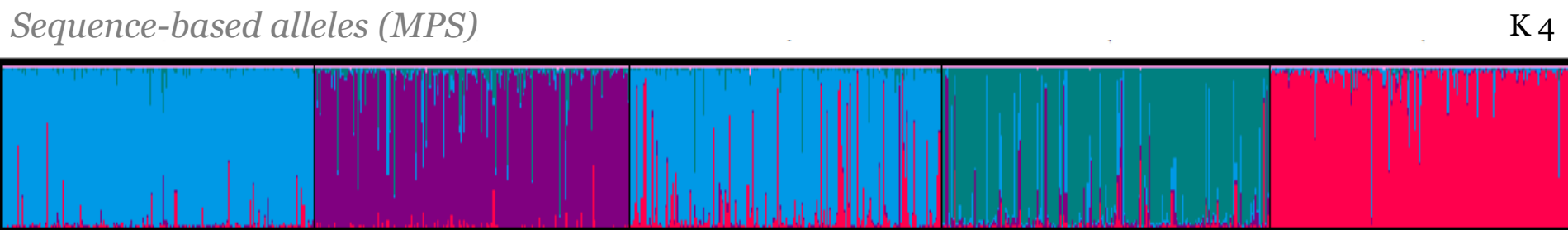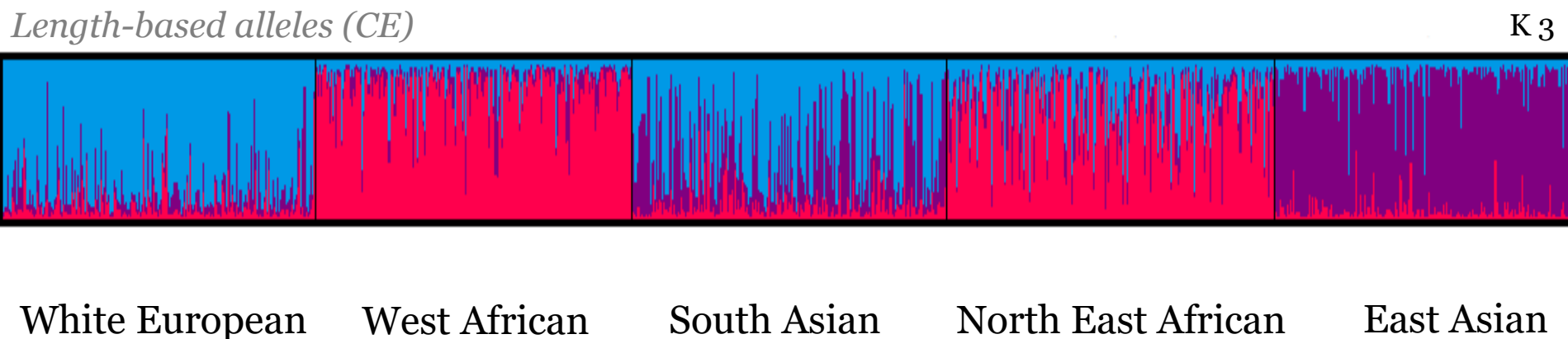
- Population differentiation using sequence-specific alleles from 27 **forensic autosomal STRs**

Using the flanking region sequences

Length based alleles (CE)

White European

South Asian

Sequence based alleles (RR)

White European

South Asian

Sequence based alleles (RR+FR)

White European

South Asian
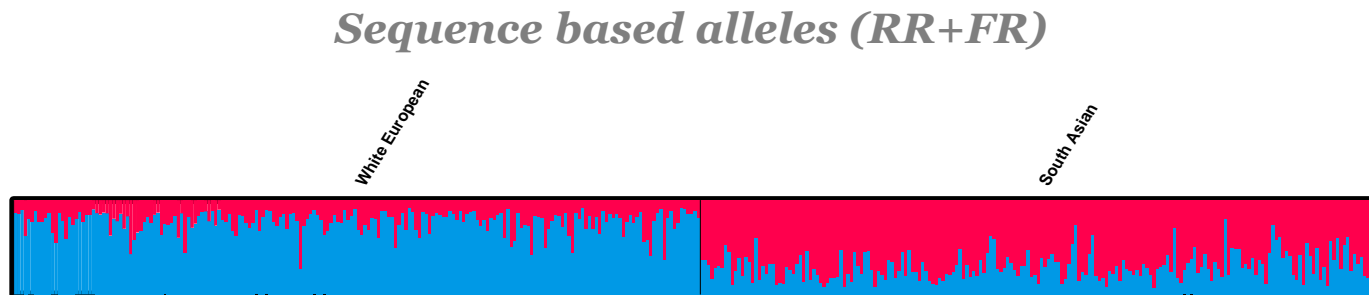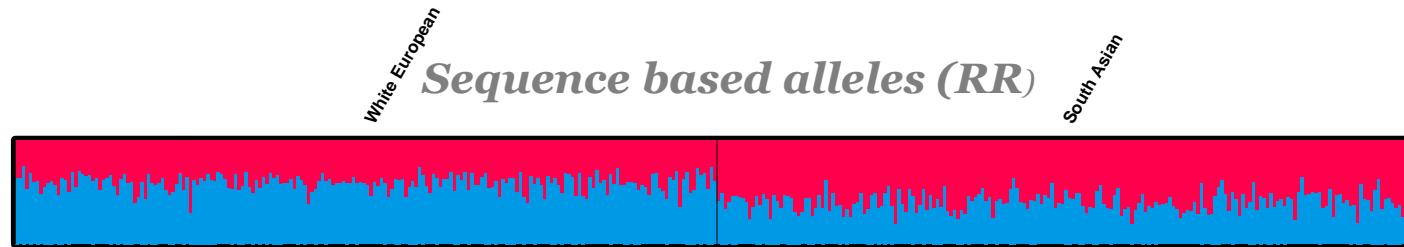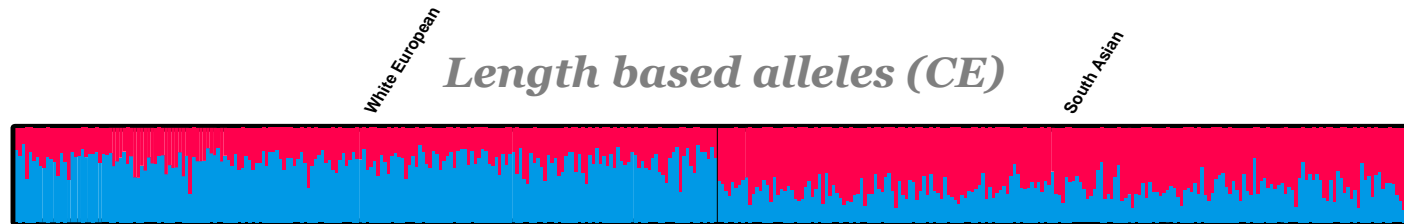
How should we report this information?

Soulbee Jin[1]* (ID)
Maretta Chase[1]*
Margaret Henry[1]
Gerry Alderson[1]
James M. Morrow[1]
Sobia Malik[1]
David Ballard[2]
Joel McGrory[1]
Neil Fernandopulle[1]
Jonathan Millman[1]
Jack Laird[1]

[1]Centre of Forensic Sciences,
 Toronto, Canada
[2]King's Forensics, King's College
 London, London, UK

Research Article

Implementing a biogeographic ancestry inference service for forensic casework

The Centre of Forensic Sciences has validated the Precision ID Ancestry Panel on the Ion S5™ Massively Parallel Sequencing instrument for use in forensic casework. The focus of this paper is the development of reporting guidelines for implementation of the biogeographic ancestry inference service based on the Admixture Prediction results produced using the Torrent Suite™ Software (Thermo Fisher Scientific). The Admixture Prediction algorithm estimates the genetic ancestry of a sample using seven root populations (Europe, East Asia, Oceania, America, Africa, South Asia, and Southwest Asia). For individuals that declared a single ancestry, there was a high correlation between the declared ancestry and the ancestry predicted by the algorithm. However, some individuals with declared ancestries of Southern Europe, Southwest Asia, South Asia and Horn of Africa had Admixture Predictions that were composed of two or more root populations at 20% or greater. For individuals with known admixed ancestry, the major component of their declaration was included in their results in all but one case. Based on these results, reporting guidelines were developed and subsequently evaluated using the Admixture Predictions of additional samples. This paper discusses the development and evaluation of these reporting guidelines, along with an implementation plan for forensic casework.

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Admixture predictions based on root populations

Root populations:
Europe, Oceania, East Asia, Africa, South Asian, America, Southwest Asia

Declared ancestry (no admixture)	Percentage concordance with prediction
Northern Europe	**80 − 100**
Southern Europe	55-95
Sub-Saharan Africa	**70-100**
North-East Africa	30-70
North Africa	8-10
East Asia	**80-100**
South Asia	35-95
Middle East	30-90

Reporting guidelines (CFS Ontario)

- **Single inclusion**
- One root population >80%
- All other populations 15% or less
- **Single mixed inclusion (single root or mixed with at least one other)**
- One root 55% - 75%
- All other populations 15% or less
- **Double inclusion (single root of either, or mixture of both)**
- Two roots 20% or greater
- **Uninformative**
- At least three roots 20% or greater

Predictions – King's Forensics – using LR

		Prediction -log(likelihood)					
		Caucasian	West African	Somalian	Chinese	Pakistan	Bangladesh
1	Caucasian	20.04	76.27	64.14	76.89	29.25	41.49
2	Caucasian	26.46	82.58	70.95	80.87	36.47	46.33
3	Caucasian	24.78	78.95	50.31	68.15	28.19	35.89
4	Caucasian	21.59	77.62	62.11	74.26	30.27	41.65
5	Caucasian	24.51	79.22	63.23	68.49	33.80	44.80
6	Caucasian	23.59	70.57	49.79	61.91	31.40	36.78
7	Caucasian	23.23	72.29	54.43	76.80	34.19	43.51
8	Caucasian	31.21	71.97	53.91	63.99	36.74	42.77
9	Caucasian	20.05	96.73	67.71	75.37	33.46	46.04
10	Caucasian	21.97	83.32	59.99	71.56	30.91	39.58
11	Caucasian	23.05	76.33	56.06	64.81	27.20	37.74
12	Caucasian	20.34	77.33	63.08	79.41	33.16	42.50
13	Caucasian	28.89	72.35	58.27	61.81	33.54	36.48
14	Caucasian	29.49	79.64	70.68	87.78	36.81	50.01
15	Caucasian	35.29	73.78	59.55	75.00	38.53	48.93
16	Caucasian	19.09	78.45	62.98	75.73	27.90	38.99
17	Caucasian	17.25	83.56	72.10	86.35	32.67	47.01

Predictions

		Prediction -log(likelihood)					
		Caucasian	West African	Somalian	Chinese	Pakistan	Bangladesh
21	West_Africa	129.59	29.08	39.05	71.59	84.32	67.87
22	West_Africa	103.78	24.57	28.61	70.72	70.82	59.88
23	West_Africa	105.33	22.64	34.00	77.97	75.89	62.81
24	West_Africa	123.72	23.96	35.07	82.72	89.09	76.34
25	West_Africa	99.70	21.27	32.26	65.46	71.97	59.74
26	West_Africa	91.74	26.65	35.08	72.01	70.68	59.43
27	West_Africa	112.01	25.02	37.69	73.69	82.36	67.61
28	West_Africa	108.80	28.38	35.62	68.26	78.25	65.91
29	West_Africa	114.97	24.47	36.72	76.64	84.48	69.92
30	Somalian	81.39	38.27	29.87	61.67	59.47	50.69
31	Somalian	94.85	43.40	26.56	54.77	62.44	54.84
32	Somalian	108.20	42.23	31.40	59.19	65.17	52.78
33	Somalian	88.27	29.63	32.55	73.85	72.00	61.84
34	Somalian	70.87	36.72	27.32	64.06	52.90	48.06
35	Somalian	89.28	46.44	34.87	61.69	57.38	48.81
36	Somalian	67.13	52.22	27.47	48.78	46.13	39.76
37	Somalian	77.53	49.74	32.59	57.76	49.62	44.32
38	Somalian	85.90	49.60	31.39	59.73	59.35	53.82

Predictions

		Prediction -log(likelihood)					
		Caucasian	West African	Somalian	Chinese	Pakistan	Bangladesh
39	Chinese	101.79	88.08	73.23	20.24	48.74	37.64
40	Chinese	89.39	91.11	71.08	29.18	48.10	41.26
41	Chinese	96.65	96.60	79.97	24.22	49.29	38.24
42	Chinese	94.42	88.14	68.97	22.75	44.35	35.05
43	Chinese	83.64	76.89	69.38	23.64	46.19	36.46
44	Chinese	96.04	91.07	76.82	18.77	43.00	36.28
45	Chinese	99.57	73.02	61.39	24.55	53.47	38.91
46	Pakistan	47.71	75.54	55.32	48.34	24.04	26.44
47	Pakistan	58.00	90.15	66.12	62.48	28.42	31.37
48	Pakistan	63.56	68.57	54.07	39.78	34.56	29.83
49	Pakistan	33.58	70.42	47.25	56.49	23.67	26.54
50	Bangladesh	60.23	62.97	50.51	43.94	32.69	22.12
51	Bangladesh	56.14	80.31	45.06	44.10	32.72	23.58
52	Bangladesh	51.34	56.49	39.65	42.62	34.20	30.67
53	Bangladesh	56.92	65.92	45.83	39.75	30.02	22.99

How good is the prediction? – blind test

Prediction Given	Caucasian	African West	African N/NE	African region unclear	South Asian	East Asian	Unclear
Caucasian	84%						15%
Nigerian/ Jamaican		50%		47%			3%
Somali			71%	23%			6%
Chinese						94%	6%
Pakistan					100%		0%
Bangladesh					90%		10%

What is the unclear?
        Mixed population within parents/grandparents
        Close population group not represented in model
STRUCTURE, Y STRs, MtDNA

Conclusions

Recommendations for phenotypic applications

- Methodologies should be validated using samples representative of the target population groups
- Support validation with blind tests
- Purpose of test must be to aid an investigation in order to assist prioritisation, and only when:
  - An STR profile, suitable for to load to a database, has revealed no hits
  - Only if sufficient of a high quality DNA extract is available

Recommendations for guidelines

- Convey the limitations of the test and mitigate the risks of misinterpretation through:
  - A statement explaining the limitations of the relevant tests to form part of the service request
  - Where possible the risk of a wrong attribution should be stated (blind tests can assist here)
  - Make it clear to the investigator that no person should be excluded from further investigation as a consequence of the test result
  - Provide a glossary to explain and clarify;:
    - Differences between geographic ancestry and where a person was born
    - Genetic ancestry and cultural ancestry (ethnicity)
    - Expected variability within the population
    - Model will depend on accuracy of declarations made by test subjects
    - Caution to be considered in admixed ancestry individuals

Thank you