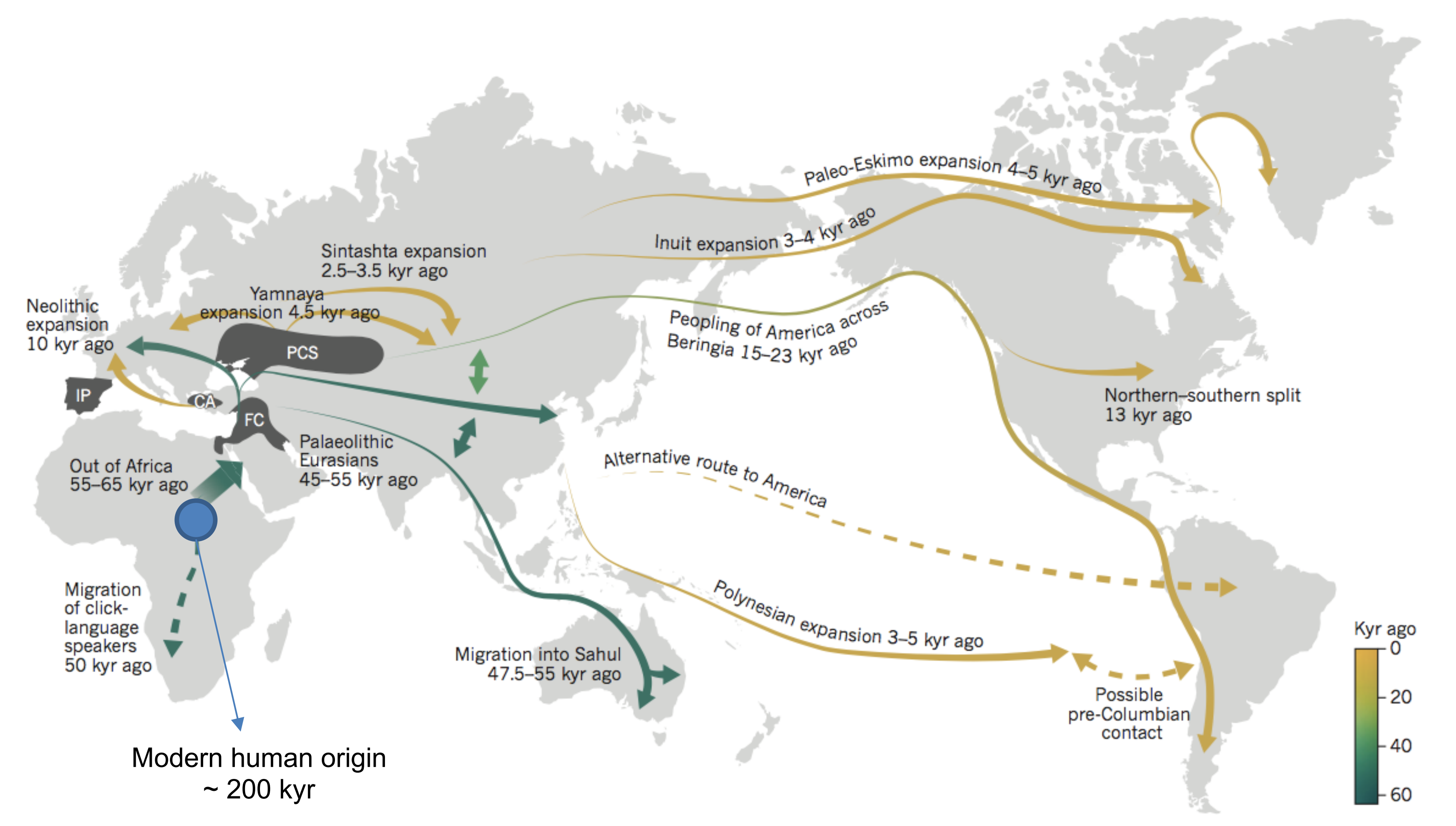
Biogeographical ancestry

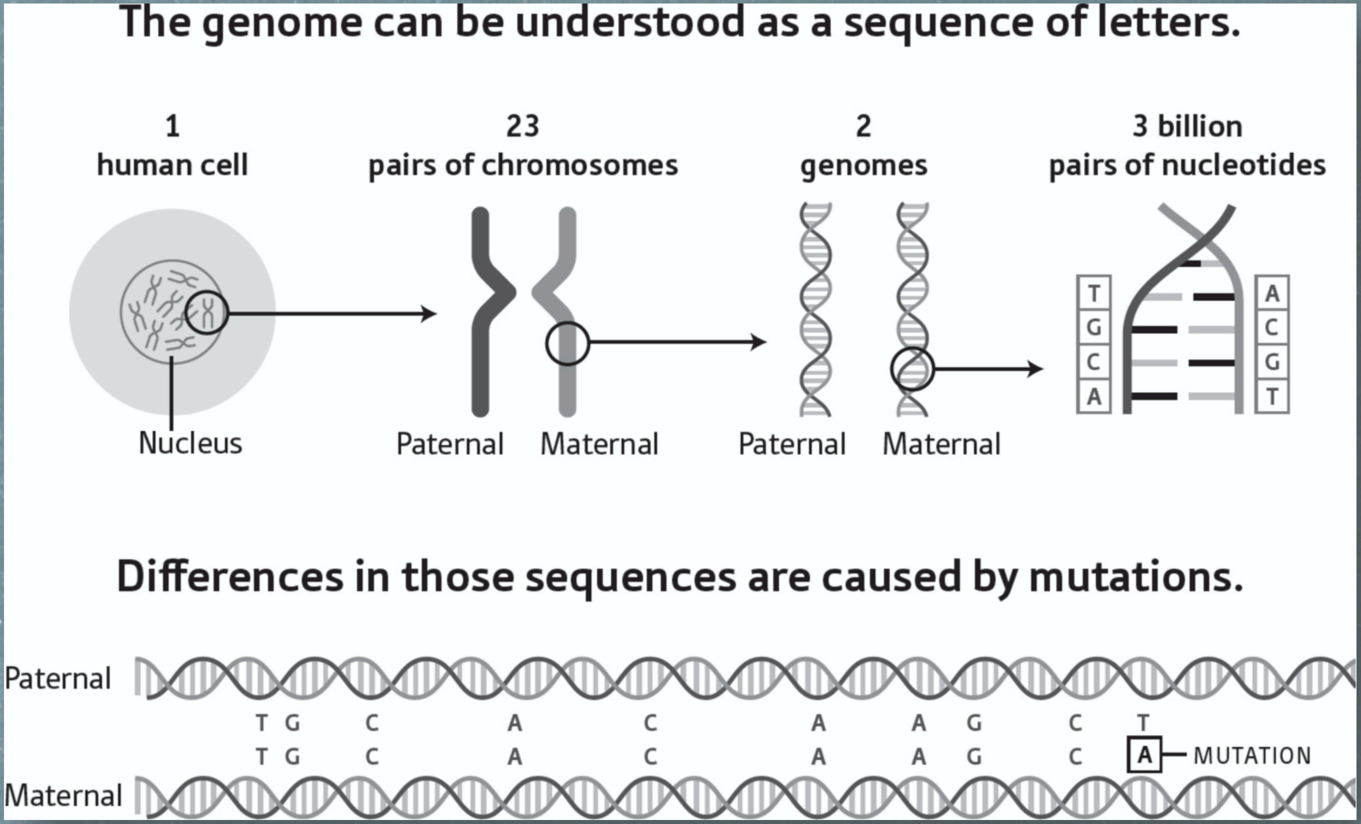
The colonizing adventure



Major human migrations across the world inferred through analyses of genomic data. Some migration routes remain under debate. For example, there is still some uncertainty regarding the migration routes used to populate the Americas. Genomic data are limited in their resolution to determine paths of migration because further population movements, subsequent to the initial migrations, may obscure the geographic patterns that can be discerned from the genomic data. Proposed routes of migration that remain controversial are indicated by dashed lines. CA, Central Anatolia; FC, Fertile Crescent; IP, Iberian Peninsula; PCS, Pontic–Caspian steppe.

Nielsen et al., Nature, 2017

How to read the DNA book



“The genome contains about three billion nucleotides, which can be thought of as four letters in a biological alphabet: adenine (A), cytosine (C), guanine (G), and thymine (T). Around 99.9 percent of these letters are identical across two lined-up genomes, but in that last ~0.1 percent there are differences, reflecting mutations that accumulate over time. These mutations tell us how closely related two people are and record exquisitely precise information about the past.”

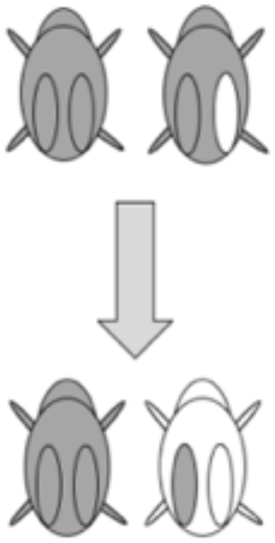
Figure from David Reich. “Who We Are and How We Got Here”.

The engines of evolution

The factors leading to changes in allele frequencies

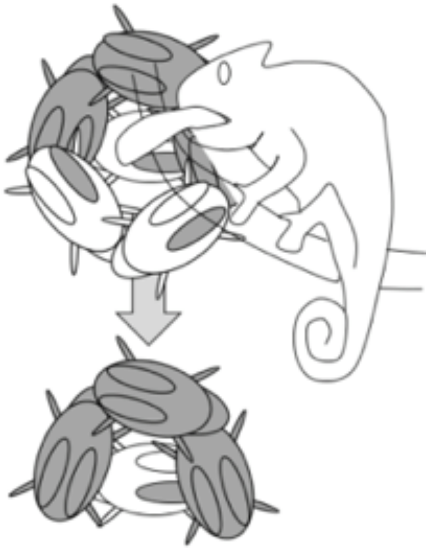
MUTATION

Creates new alleles in a gene pool



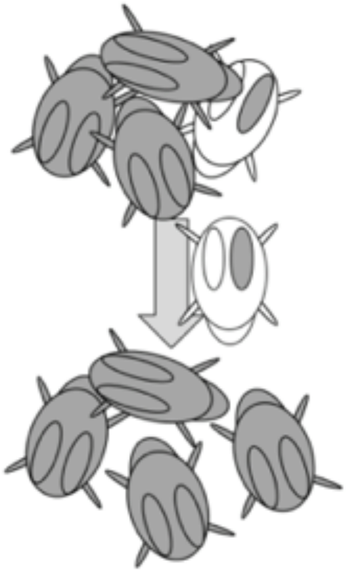
SELECTION

Can be a major force driving allele frequency change, and leads to adaptation



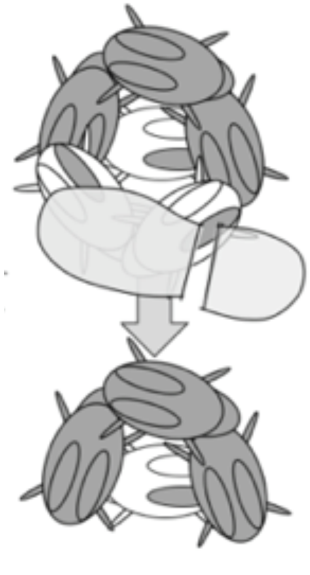
MIGRATION

Gene flow from other populations can alter allele frequencies

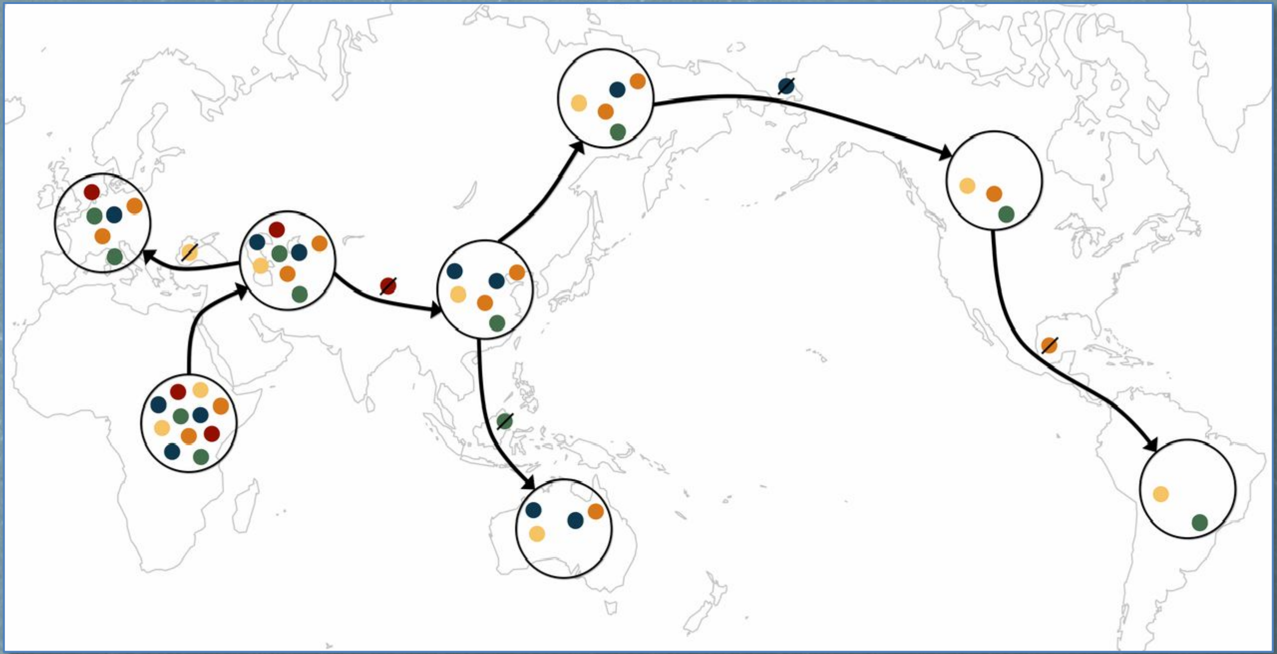


GENETIC DRIFT

Causes random changes in allele frequency especially in small populations



African origins



The serial founder model in human evolution. (A) A schematic of the model. Each color represents a distinct allele. Migration events outward from Africa tend to carry with them only a subset of the genetic diversity from the source population, and some alleles are lost during migration events.

Noah A. Rosenberg and Jonathan T. L. Kang, 2015, GENETICS

Genetic variation decreases moving farther away from Africa

SERIAL FOUNDER EFFECT

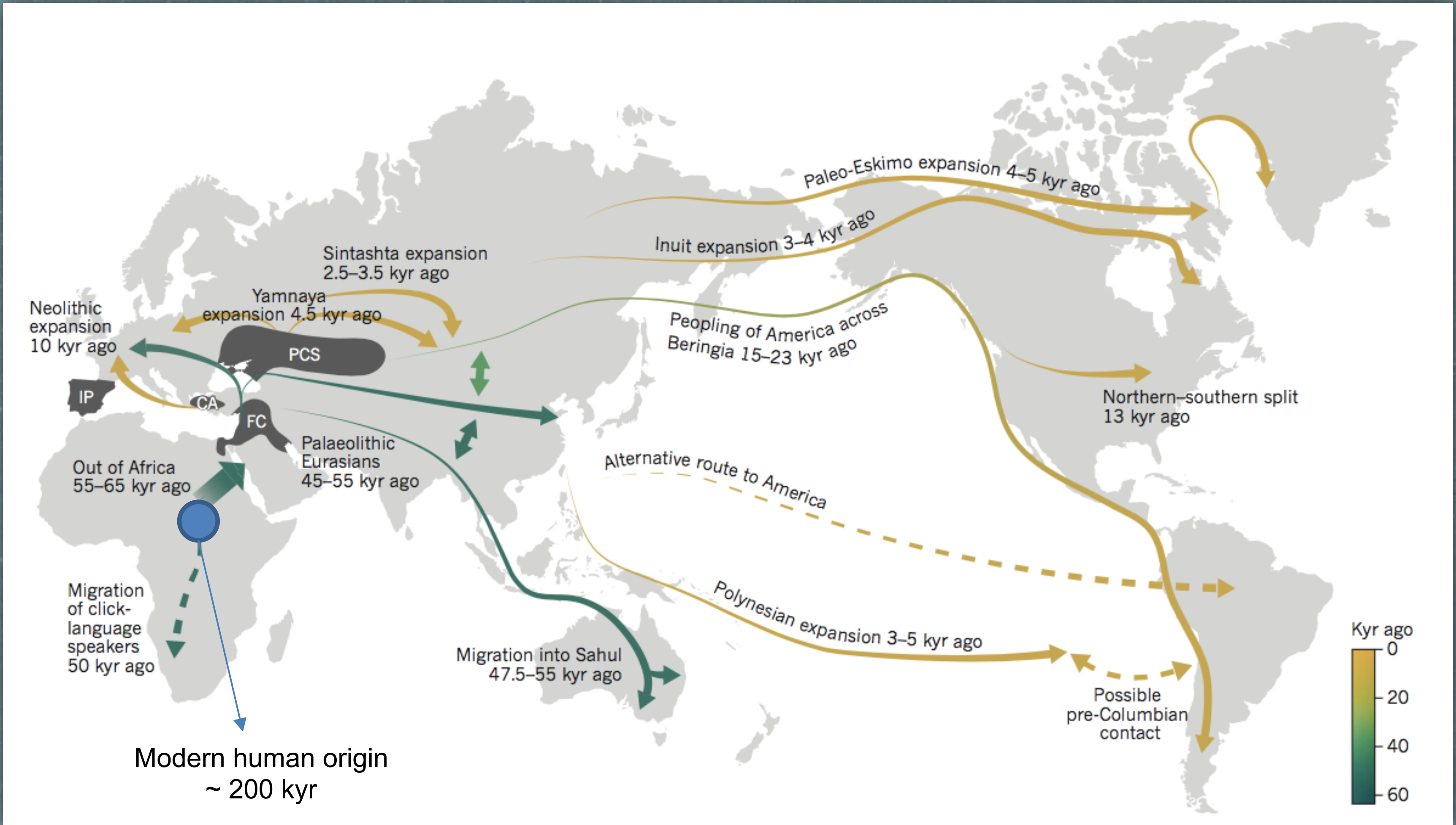
Populations at increasing distance from Africa show more genetic differences from the African populations

ISOLATION BY DISTANCE

Africans show the most shuffling of genomic segments of all human populations

PATTERNS OF LINKAGE DISEQUILIBRIUM

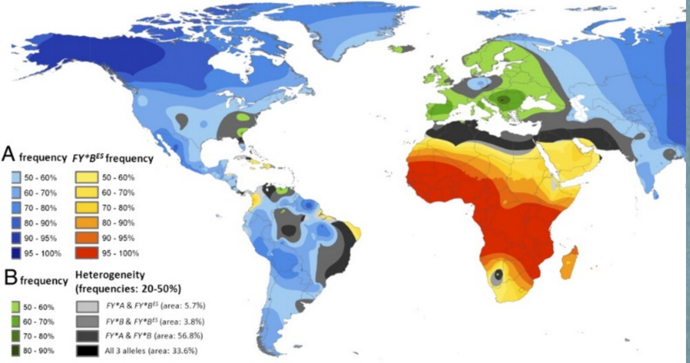
The colonizing adventure

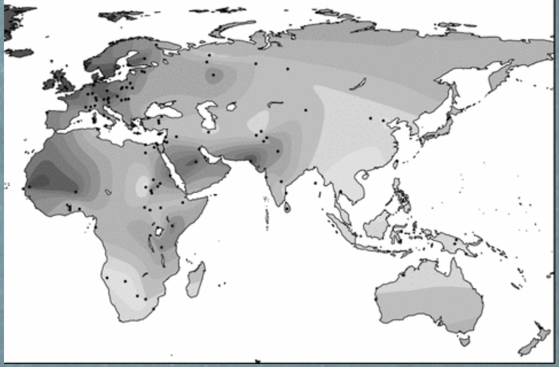


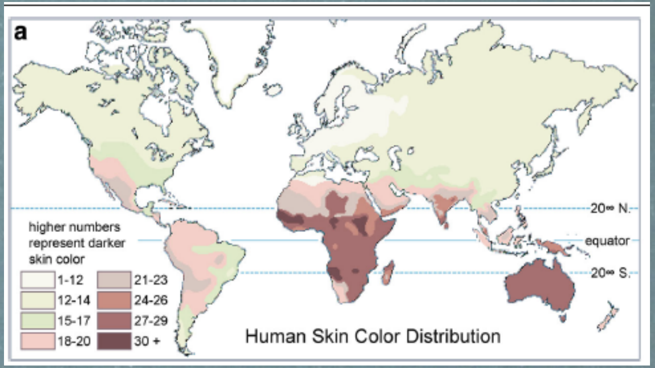
Major human migrations across the world inferred through analyses of genomic data. Some migration routes remain under debate. For example, there is still some uncertainty regarding the migration routes used to populate the Americas. Genomic data are limited in their resolution to determine paths of migration because further population movements, subsequent to the initial migrations, may obscure the geographic patterns that can be discerned from the genomic data. Proposed routes of migration that remain controversial are indicated by dashed lines. CA, Central Anatolia; FC, Fertile Crescent; IP, Iberian Peninsula; PCS, Pontic–Caspian steppe.

Nielsen et al., Nature, 2017

- ✓ Occasionally, genetic innovations that occurred locally were environmentally induced or modified






a
higher numbers represent darker skin color
1-12
12-14
15-17
18-20
21-23
24-26
27-29
30 +
20∞ N.
equator
20∞ S.
Human Skin Color Distribution

Disease

T>C mutation at rs2414778 leads to failure of Duffy antigen expression on the surface of red blood cells in humans, conferring resistance to Plasmodium vivax malaria

Diet

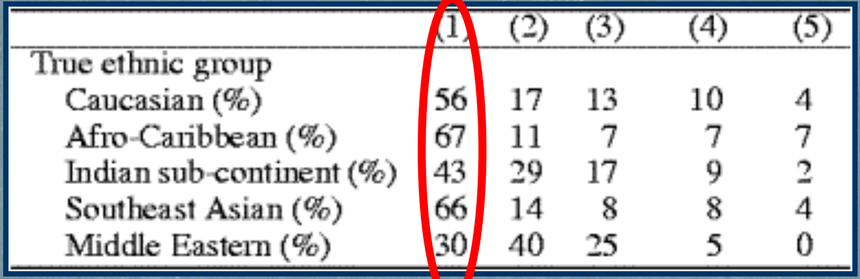
Lactase production persistence is subject to strong positive selection in the population with cultural practice of milking

Climate

In tropical (original human) environment, dark skin protects photodegradation of cutaneous and systemic folate.
N or S of 43° dark skin leads to vitamin D deficiency

Forensic STRs as Ancestry Informative Markers (AIMs)

High mutation rate
Not under selection
Not originally chosen for ancestry inference
Originally chosen to be highly polymorphic in all populations

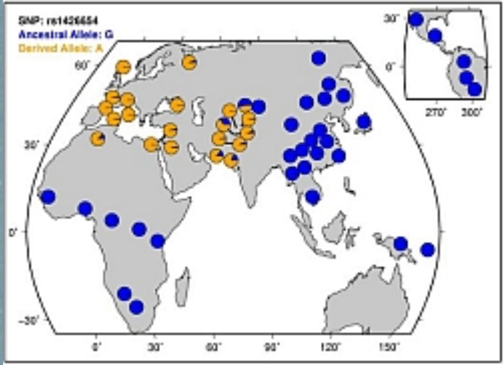


	(1)	(2)	(3)	(4)	(5)
True ethnic group					
Caucasian (%)	56	17	13	10	4
Afro-Caribbean (%)	67	11	7	7	7
Indian sub-continent (%)	43	29	17	9	2
Southeast Asian (%)	66	14	8	8	4
Middle Eastern (%)	30	40	25	5	0

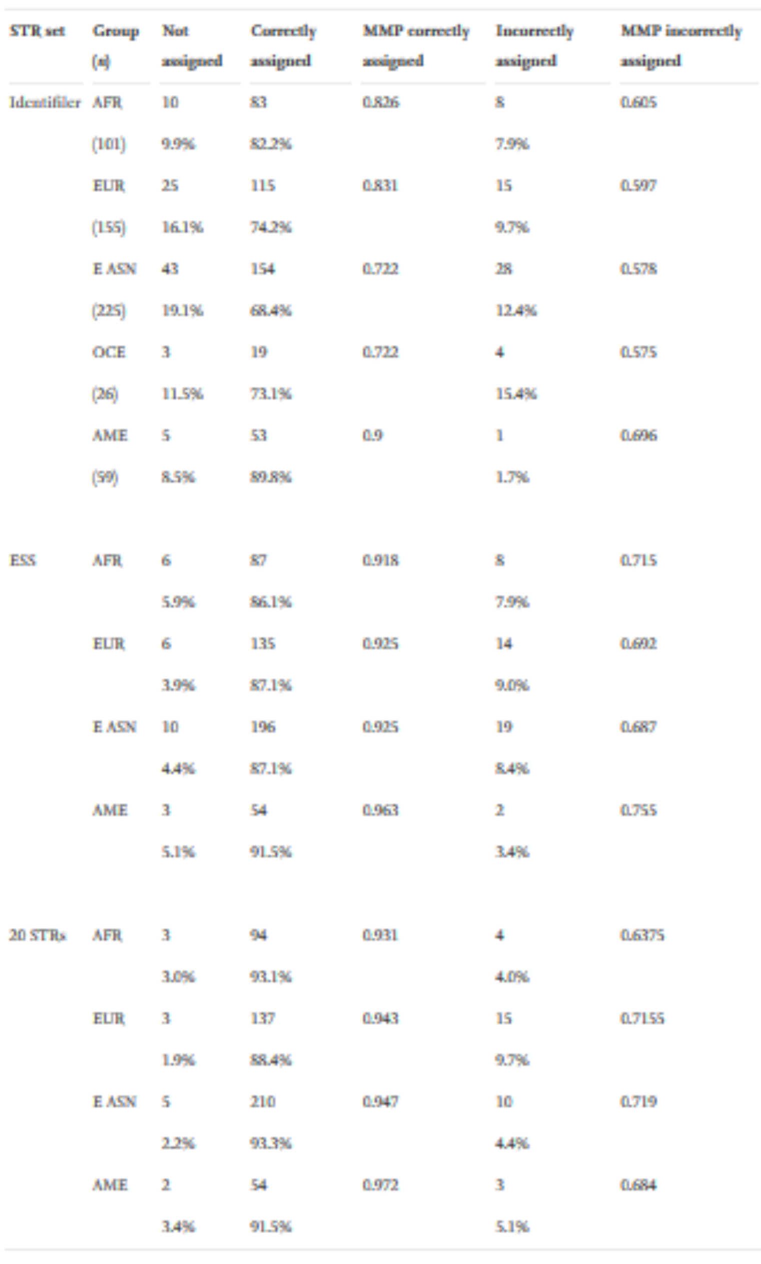
Lowe et al Forensic Sci Int 2001

UK criminal DNA database
6 SGM STR loci

Correctly assigned (%)



Human Genome Diversity Project (HGDP) populations



STR set	Group (n)	Not assigned	Correctly assigned	MMP correctly assigned	Incorrectly assigned	MMP incorrectly assigned
Identifiler	AFR	10	83	0.826	8	0.605
	(101)	9.9%	82.2%		7.9%	
	EUR	25	115	0.831	15	0.597
	(155)	16.1%	74.2%		9.7%	
	E ASN	43	154	0.722	28	0.578
	(225)	19.1%	68.4%		12.4%	
	OCE	3	19	0.722	4	0.575
	(26)	11.5%	73.1%		15.4%	
	AME	5	53	0.9	1	0.696
	(59)	8.5%	89.8%		1.7%	
ESS	AFR	6	87	0.918	8	0.715
		5.9%	86.1%		7.9%	
	EUR	6	135	0.925	14	0.692
		3.9%	87.1%		9.0%	
	E ASN	10	196	0.925	19	0.687
		4.4%	87.1%		8.4%	
	AME	3	54	0.963	2	0.755
		5.1%	91.5%		3.4%	
20 STRs	AFR	3	94	0.931	4	0.6375
		3.0%	93.1%		4.0%	
	EUR	3	137	0.943	15	0.7155
		1.9%	88.4%		9.7%	
	E ASN	5	210	0.947	10	0.719
		2.2%	93.3%		4.4%	
	AME	2	54	0.972	3	0.684
		3.4%	91.5%		5.1%	

Phillips et al Forensic Sci Int Genet 2011

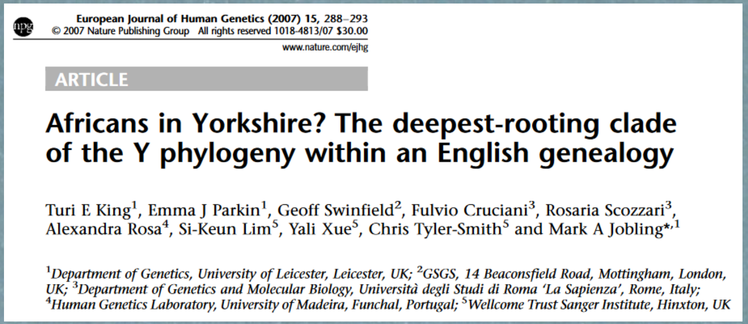
Forensic Y-STRs and mtDNA control region as AIMs

✓ highly differentiated geographically, because of uniparental inheritance and reduced effective population size

✓ STR / control region haplotypes can be used (at a certain extent) to derive SNP variation (in the coding region, for mtDNA): «Haplogroups» which are more evolutionary stable

✓ reflect distant male/female lineages that can have little to do with our current appurtenance, and can therefore be misleading as a genetic-base substitute for eyewitness testimony

✓ the problem is particularly evident in recently admixed populations, but it can occur anywhere…



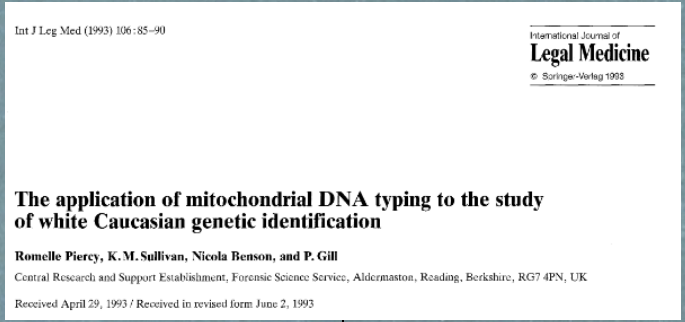
European Journal of Human Genetics (2007) 15, 288–293
© 2007 Nature Publishing Group All rights reserved 1018-4813/07 $30.00
www.nature.com/ejhg

ARTICLE

Africans in Yorkshire? The deepest-rooting clade of the Y phylogeny within an English genealogy

Turi E King[1], Emma J Parkin[1], Geoff Swinfield[2], Fulvio Cruciani[3], Rosaria Scozzari[3], Alexandra Rosa[4], Si-Keun Lim[5], Yali Xue[5], Chris Tyler-Smith[5] and Mark A Jobling*,[1]

[1]Department of Genetics, University of Leicester, Leicester, UK; [2]GSGS, 14 Beaconsfield Road, Mottingham, London, UK; [3]Department of Genetics and Molecular Biology, Università degli Studi di Roma 'La Sapienza', Rome, Italy; [4]Human Genetics Laboratory, University of Madeira, Funchal, Portugal; [5]Wellcome Trust Sanger Institute, Hinxton, UK



Int J Leg Med (1993) 106 : 85–90

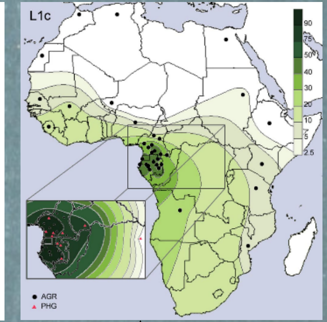
International Journal of
Legal Medicine
© Springer-Verlag 1993

The application of mitochondrial DNA typing to the study of white Caucasian genetic identification

Romelle Piercy, K. M. Sullivan, Nicola Benson, and P. Gill

Central Research and Support Establishment, Forensic Science Service, Aldermaston, Reading, Berkshire, RG7 4PN, UK

Received April 29, 1993 / Received in revised form June 2, 1993



Description of one individual displaying a typically sub-Saharan African mtDNA type

Specifically selected AIMs
(not part of standard panels of identificative markers)

✓ How diverse are we?

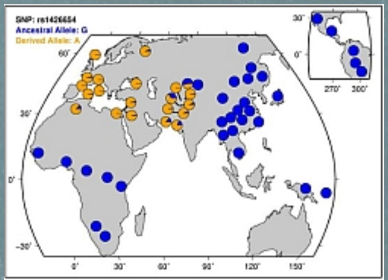
- Analysis of MOlecular VAriance (AMOVA): how the observed differences in allele frequencies are explained by

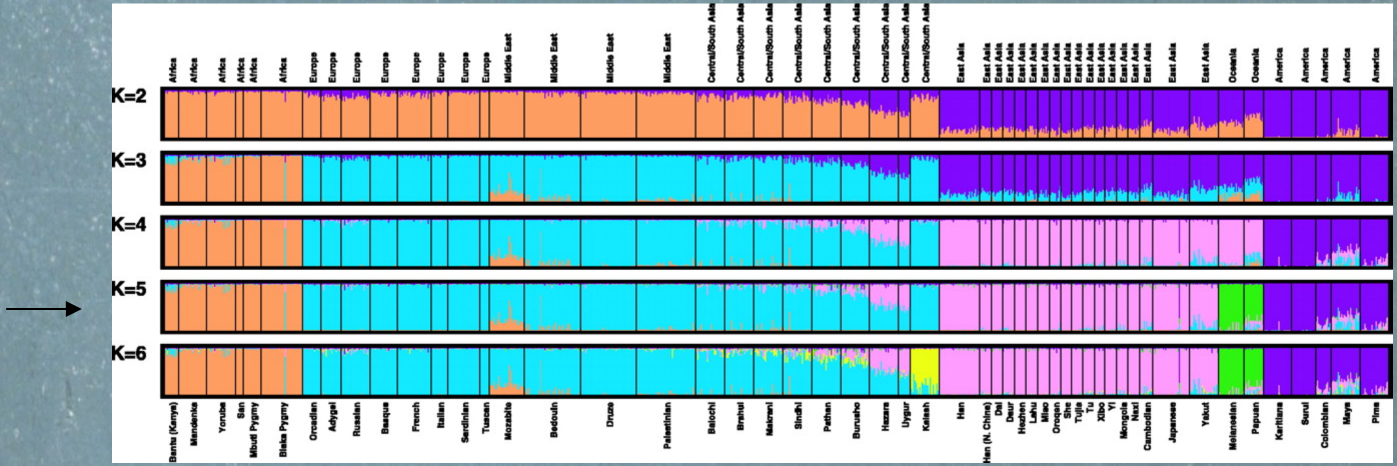
- Differences between groups of populations
- Differences between populations within groups
- Differences between individuals within populations

↓

Blood groups and proteins (Lewontin Evol Biol 1972) 85%
STRs (Rosenberg et al. Science 2002) 93%
SNPs (Kidd et al J Heredity 2004) 86% → Better detectors of population structure

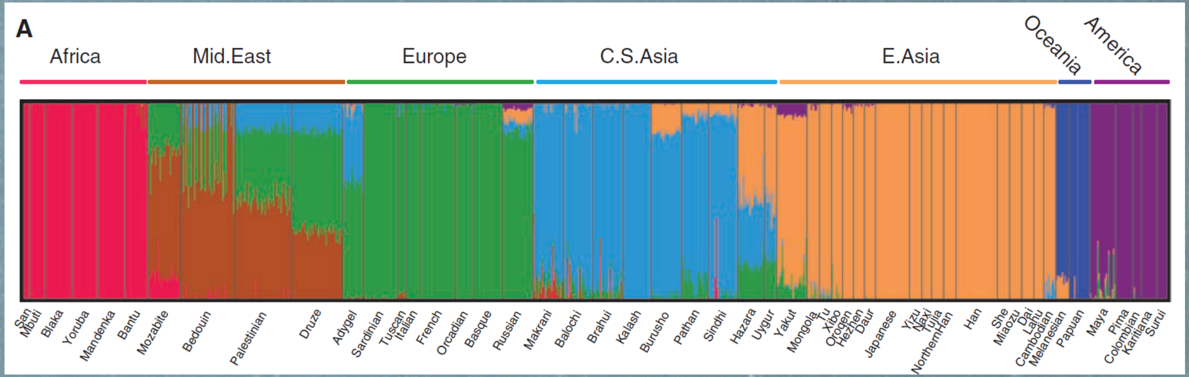
- ✓ Is it possible to use genetic data backwards (group assignment from DNA data)?
- A set of worldwide population samples (HGDP) is considered
- a genetic clustering algorithm (STRUCTURE analysis) is used to define, based on each individual’s similarity or dissimilarity to others in the sample set, the cluster number (k) with maximum likelihood





377 STRs: optimal value of K=5 (Eurasia, sub-Saharan Africa, East Asia, America and Oceania) (Rosenberg et al Science 2002)

K7



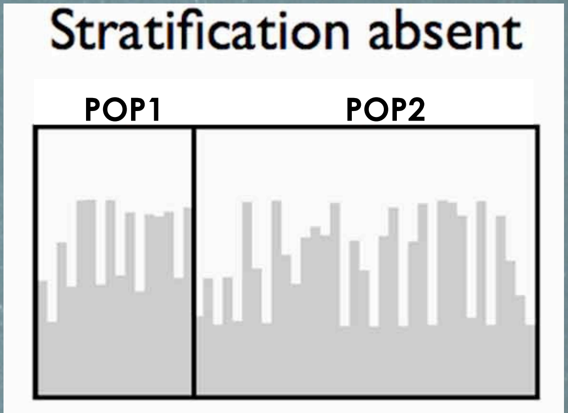
650,000 SNPs: at K=7 a sixth component typical of the Indian subcontinent is clearly identified; the seventh component occurs at higher proportion in the Middle East(though far from 100%) (Li et al Science 2008)

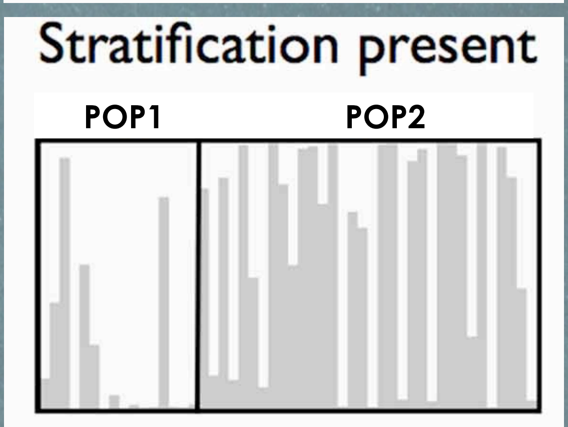
✓ STRUCTURE analyses the differences in the distribution of genetic variants amongst populations with a Bayesian iterative algorithm by placing samples into groups whose members share similar patterns of variation (loci are assumed to be in Hardy-Weinberg equilibrium and independent, i.e. not in linkage disequilibrium).

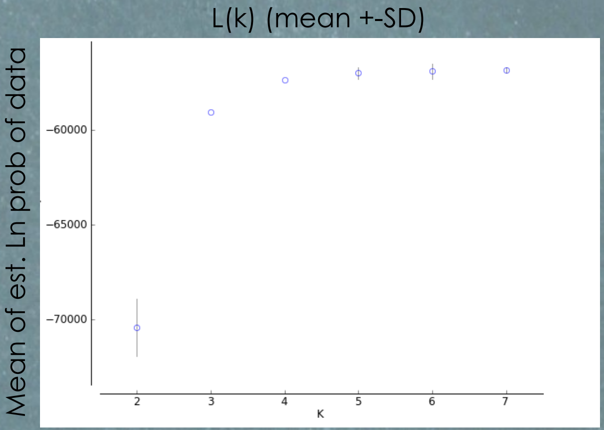
✓ STRUCTURE uses a systematic Bayesian clustering approach applying Markov Chain Monte Carlo (MCMC) estimation. The MCMC process begins by randomly assigning individuals to a pre-determined number of groups (K), then variant frequencies are estimated in each group and individuals re-assigned based on those frequency estimates. This is repeated many times, resulting in a progressive convergence toward reliable allele frequency and estimates in each population and membership probabilities of individuals to a population.

✓ STRUCTURE also calculates the likelihood of the data for a range of K values (posterior probabilities). Plots of posterior probability values typically progress to a plateau for levels of K beyond the most applicable number of detected populations. So the smallest stable K value represents the optimum K value.

✓ STRUCTURE can be used two-ways: to identify populations from the data and to assign individuals to that population representing the best fit for the variation patterns found.


Stratification absent


Stratification present

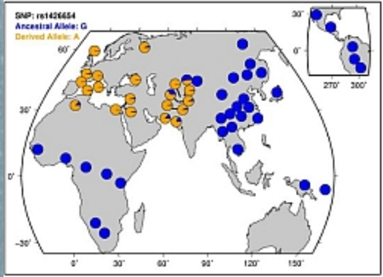


Optimal K=4

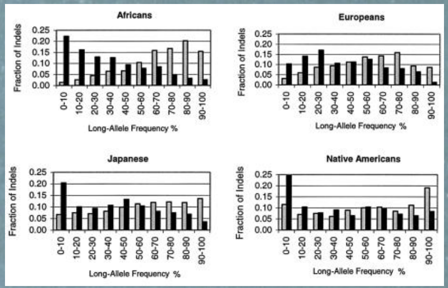
- ✓ DNA amount in forensic stains is limited
- need for compact AIMS assays
- no more than ~ 50 Indels or ~ 30 SNPs can be squeezed in a multiplex PCR / Single Base Extension reaction

↓

Selection of most informative markers starting from worldwide reference datasets



HGDP
>1000 individuals from 56 populations now typed for >1 million SNPs



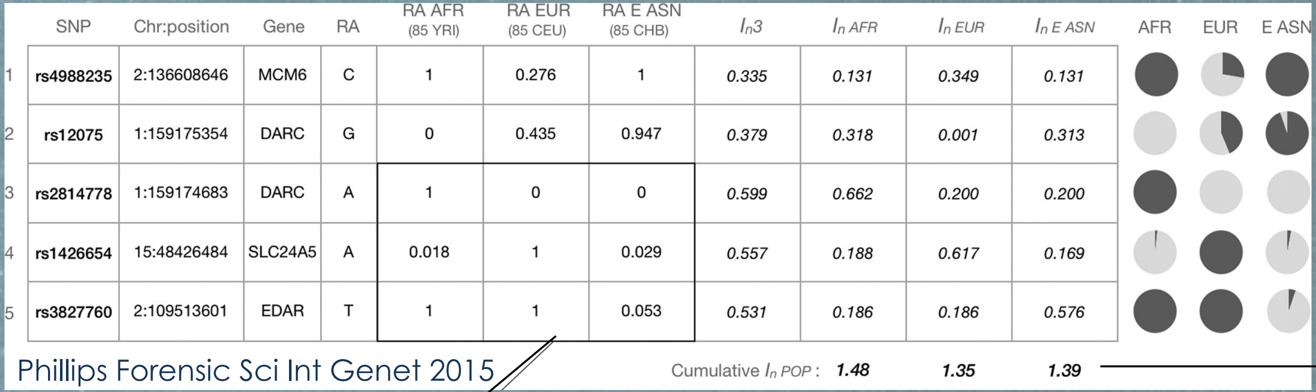
Marshfield Diallelic Insertion/ Deletion Polymorphisms database
>200 individuals (part HGDP) typed for 2,000 indels (Weber et al Am J Hum Genet 2002)

Greenland
Iceland
Finland
Sweden
Norway
Russia
United Kingdom
Poland
Germany
Ukraine
Kazakhstan
Mongolia
Canada
United States
North Atlantic Ocean
Turkey
China
South Korea
Japan
Afghanistan
Iraq
Iran
Pakistan
Algeria
Libya
Egypt
Saudi Arabia
India
Mexico
Mali
Republic of Niger
Sudan
Venezuela
Ethiopia
DR Congo
Tanzania
Indonesia
Papua New Guinea
Brazil
Angola
Bolivia
Namibia
Botswana
Madagascar
Indian Ocean
South Pacific Ocean
South Atlantic Ocean
Chile
South Africa
Australia
Argentina
New Zealand

1000 Genomes project
Whole genome sequences of >2500 individuals (Sudmant et al Nature 2015)

- ✓ Measures of divergence between populations
- F_{ST}
- δ, allele frequency differential: the absolute value of p_X-p_Y (comparing allele frequency p in populations X and Y)
- informativeness-for-assignment metric I_n (Rosenberg et al Am J Hum Genet 2003)

- ✓ Combination of selected markers in order to achieve equal capacity to differentiate between target population groups
- distribution of human diversity has led to strong divergence between African and other populations
- A simple list of top markers in terms of F_{ST}, δ or I_n will lead to optimal differentiation of Africans but suboptimal differentiation between non-African populations



	SNP	Chr:position	Gene	RA	RA AFR (85 YRI)	RA EUR (85 CEU)	RA E ASN (85 CHB)	I_n3	I_n AFR	I_n EUR	I_n E ASN	AFR	EUR	E ASN
1	**rs4988235**	2:136608646	MCM6	C	1	0.276	1	*0.335*	*0.131*	*0.349*	*0.131*	●	◔	●
2	**rs12075**	1:159175354	DARC	G	0	0.435	0.947	*0.379*	*0.318*	*0.001*	*0.313*	○	◑	●
3	**rs2814778**	1:159174683	DARC	A	1	0	0	*0.599*	*0.662*	*0.200*	*0.200*	●	○	○
4	**rs1426654**	15:48426484	SLC24A5	A	0.018	1	0.029	*0.557*	*0.188*	*0.617*	*0.169*	○	●	○
5	**rs3827760**	2:109513601	EDAR	T	1	1	0.053	*0.531*	*0.186*	*0.186*	*0.576*	●	●	○
								Cumulative I_n POP :	***1.48***	***1.35***	***1.39***			

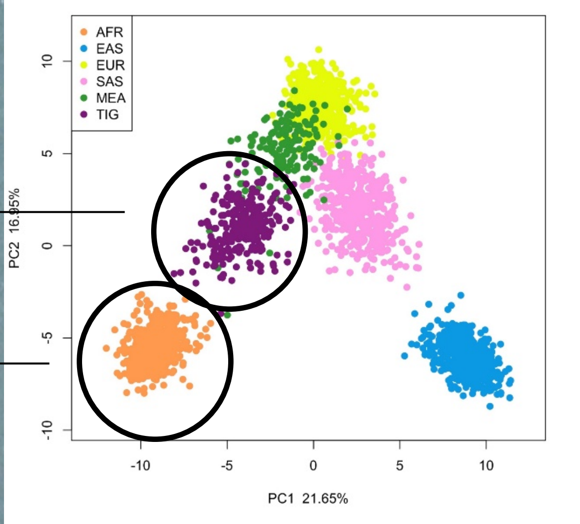
Phillips Forensic Sci Int Genet 2015

Best markers are those near fixation with a private allele specific for a single target group (very rare)

Hard to preserve a balance in population specific divergence with growing numbers of AIMs

✓ Several SNP/Indel AIMs panels have been described in recent years (see Phillips Forensic Sci Int Genet 2015 provided as supplementary material to this lesson as a general review)

- Basic differentiation is at K=3 (sub-Saharan Africa, Europe, Far East)
- most panels enabling discrimination at K=5 (sub-Saharan Africa, Europe, Far East + America and Oceania)
- Some panels allowing further discrimination within Asia (Indian sub-continent vs Far East) or Africa



Eastern Africans (Ethiopia)

Western Africans + Bantu kenyans (1000 genomes)

Kumar et al, Forensic Sci Int Genet 2020

Principal Component Analysis (PCA) is a multi-dimensional scaling (MDS) technique that reduces the dimensionality of data while keeping the largest possible portionof it variability. It provides an intuitive and simply understood way to interpret patterns of divergence amongst sets of populations.

- ✓ Final assignment of the unknown donor of a stain to a specific population group based on AIMs data can be done through Bayes analysis
- the combined genotype frequencies estimated for each population are used to calculate their likelihood. A probability of ancestry is then assigned from the ratio of the two highest likelihoods.

AIM1	AFR	EUR	ASN	AIM2	AFR	EUR	ASN	AIM3	AFR	EUR	ASN
A	.999	.001	.001	A	.018	.999	.029	A	.001	.001	.947
T	.001	.999	.999	T	.982	.001	.971	T	.999	.999	.053

Unknown

AIM1	AIM2	AIM3	pAFR	pEUR	pASN
TT	TT	AA	1×10^{-12}	1×10^{-12}	0.844

Sample classified as Asian
LR = pASN / 2° best p
$= 0.844/1\text{x}10^{-12}$
$= 8.4 \times 10^{11}$

LR		≥1			$\geq 10^1$			$\geq 10^2$			$\geq 10^3$	
	Not Classified (N)	Correctly Classified (C)	Wrongly classified (W)	N	C	W	N	C	W	N	C	W
Africans	0	.992	.008	.024	.976	.000	.079	.921	.000	.238	.762	.000
South Asians	0	.980	.020	.045	.947	.008	.092	.906	.002	.184	.816	.000
Middle Easterns	0	.822	.178	.202	.694	.104	.448	.497	.055	.656	.326	.018
Ethiopians	0	.976	.024	.075	.909	.016	.206	.790	.004	.294	.702	.004

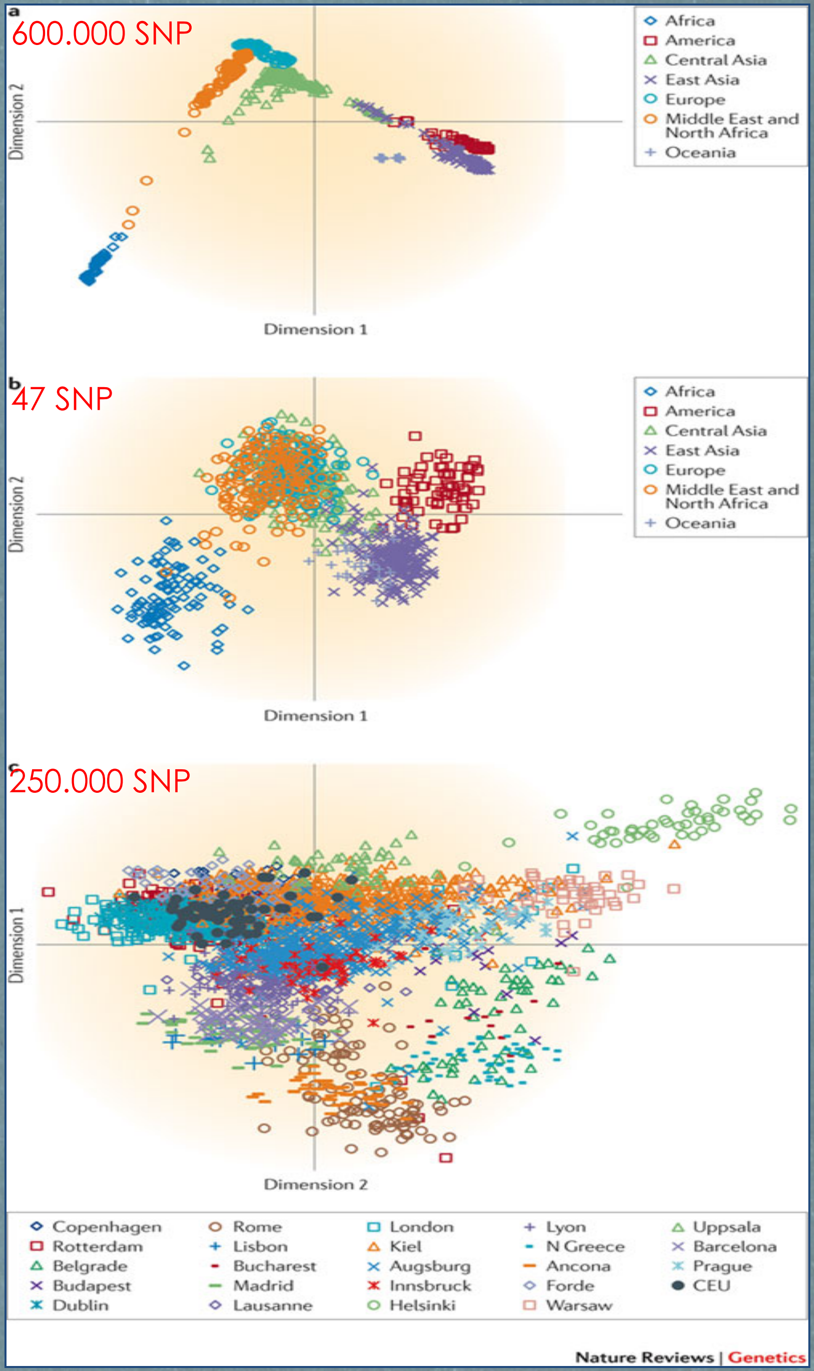
31 SNPs (Kumar et al, Forensic Sci Int Genet 2020)

LR cautionary threshold can be assigned to reduce error rates

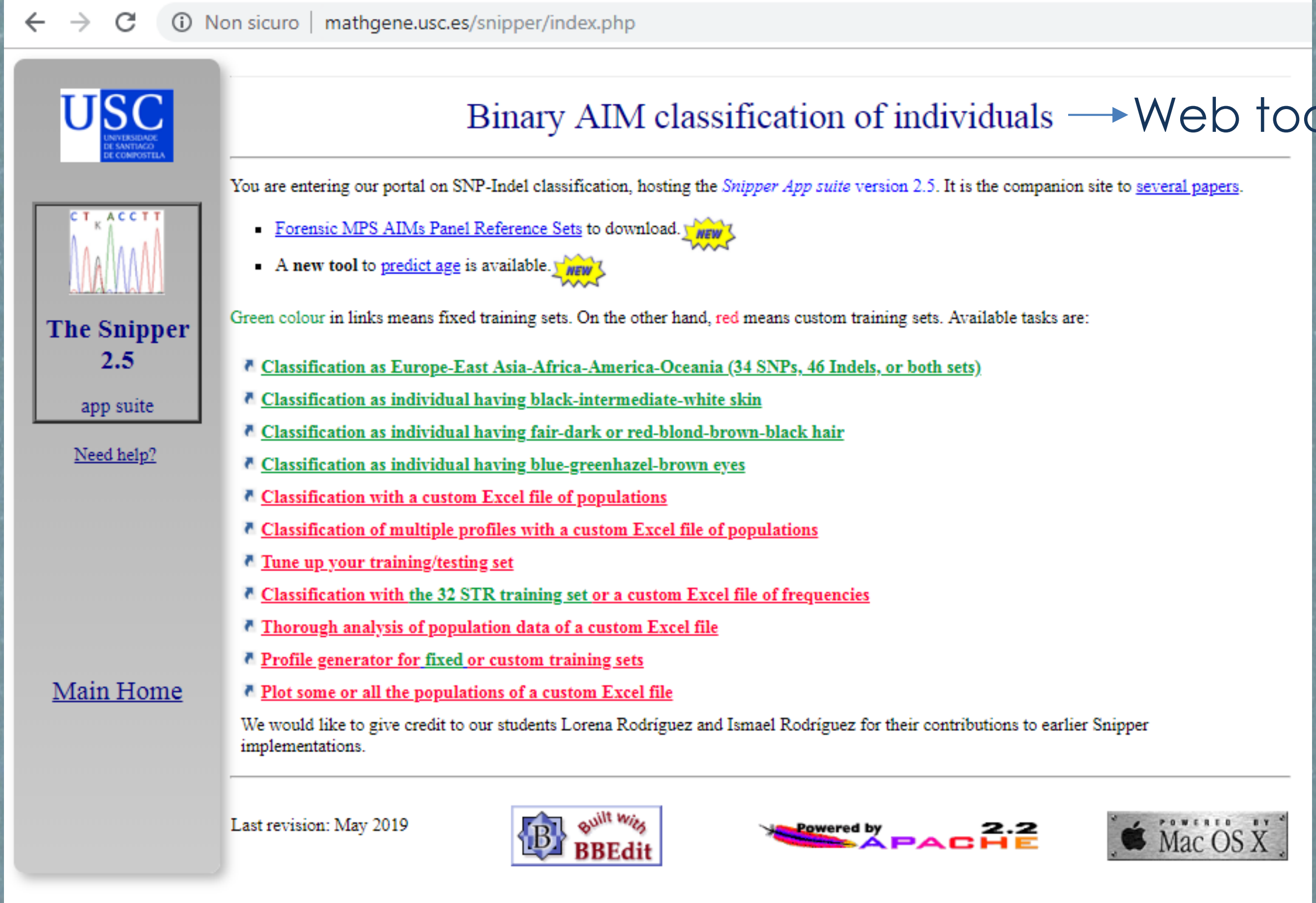
✓ How accurate can we be in our inference?

- with a limited number of SNPs (amenable for forensic investigations) it is possible to distinguish between populations at continental level (some subcontinental discrimination in Asia, Africa)

- even with hundred thousands of SNPs hard to identify geographic origin at subcontinental level in Europe



Kayser M et al. Nature Rev 2012



Binary AIM classification of individuals

You are entering our portal on SNP-Indel classification, hosting the *Snipper App suite* version 2.5. It is the companion site to several papers.

- Forensic MPS AIMs Panel Reference Sets to download.
- A **new tool** to predict age is available.

Green colour in links means fixed training sets. On the other hand, red means custom training sets. Available tasks are:

- **Classification as Europe-East Asia-Africa-America-Oceania (34 SNPs, 46 Indels, or both sets)**
- **Classification as individual having black-intermediate-white skin**
- **Classification as individual having fair-dark or red-blond-brown-black hair**
- **Classification as individual having blue-greenhazel-brown eyes**
- **Classification with a custom Excel file of populations**
- **Classification of multiple profiles with a custom Excel file of populations**
- **Tune up your training/testing set**
- **Classification with the 32 STR training set or a custom Excel file of frequencies**
- **Thorough analysis of population data of a custom Excel file**
- **Profile generator for fixed or custom training sets**
- **Plot some or all the populations of a custom Excel file**

We would like to give credit to our students Lorena Rodríguez and Ismael Rodríguez for their contributions to earlier Snipper implementations.

Last revision: May 2019

→ Web tools for ancestry inference

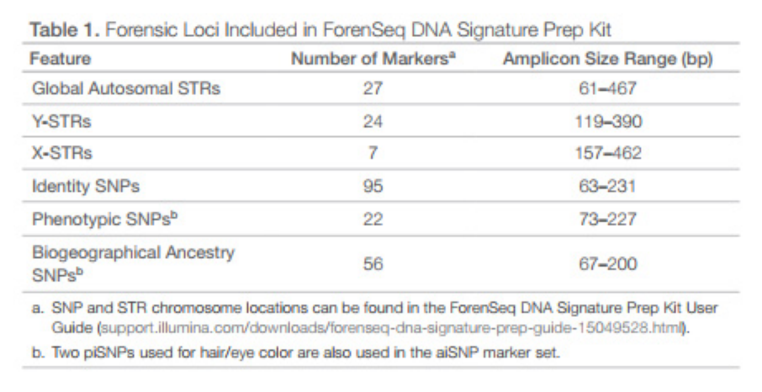


Table 1. Forensic Loci Included in ForenSeq DNA Signature Prep Kit

Feature	Number of Markers[a]	Amplicon Size Range (bp)
Global Autosomal STRs	27	61–467
Y-STRs	24	119–390
X-STRs	7	157–462
Identity SNPs	95	63–231
Phenotypic SNPs[b]	22	73–227
Biogeographical Ancestry SNPs[b]	56	67–200

a. SNP and STR chromosome locations can be found in the ForenSeq DNA Signature Prep Kit User Guide (support.illumina.com/downloads/forenseq-dna-signature-prep-guide-15049528.html).

b. Two piSNPs used for hair/eye color are also used in the aiSNP marker set.

Ancestry informative SNPs
readily combined in new MPS assays ←