POPULATION STRUCTURE

Human Populations: History and Structure

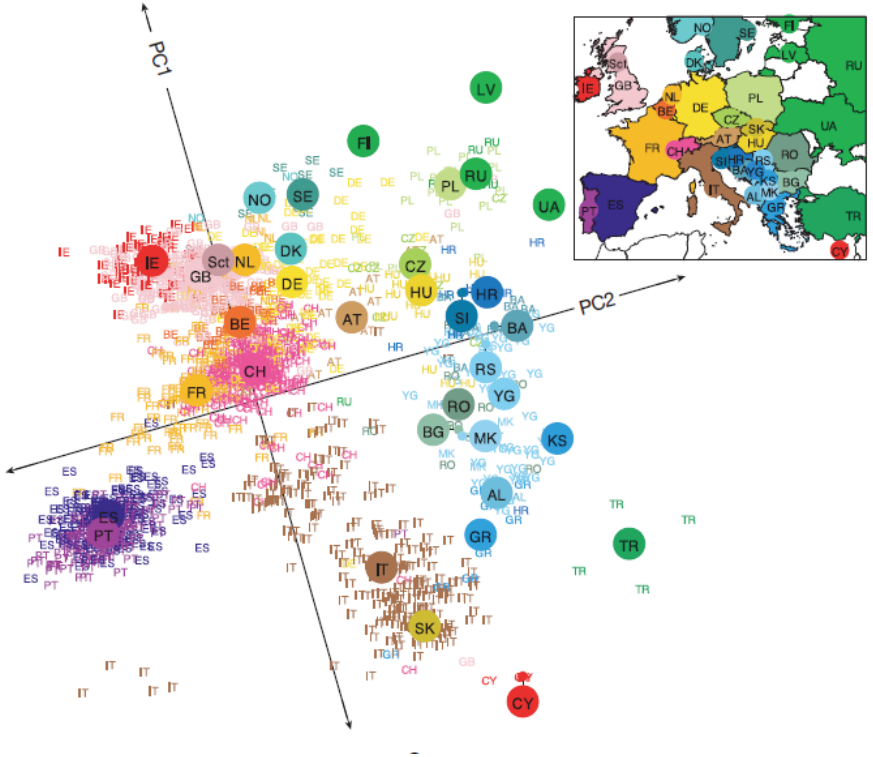
In the paper

> Novembre J, Johnson, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann A, Nelson MB, Stephens M, Bustamante CD. 2008. Genes mirror geography within Europe. Nature 456:98

there is quite dramatic evidence that our genetic profiles contain information about where we live, suggesting that these profiles reflect the history of our populations.

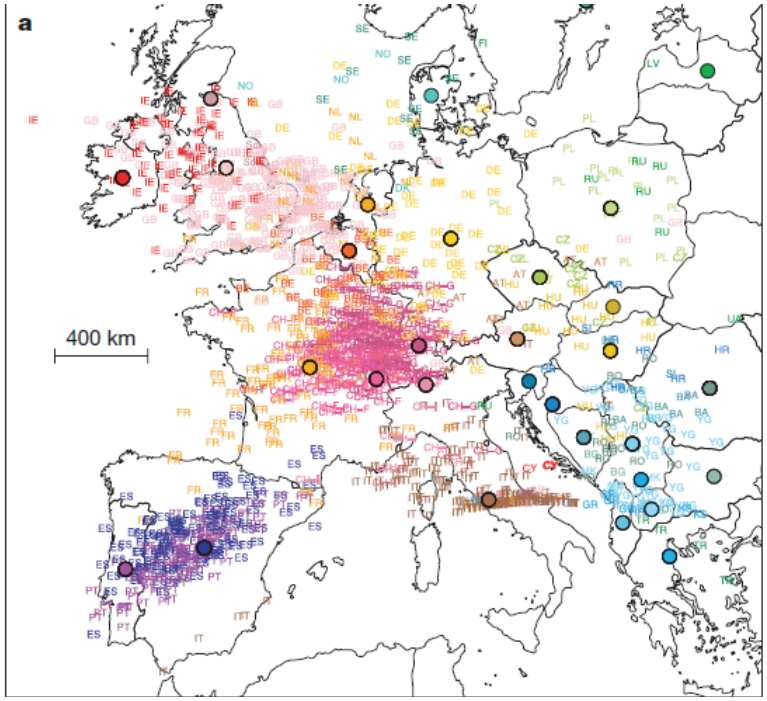
The authors collected "SNP" (single nucleotide polymorphism) data on over people living in Europe. Either the country of origin of the people's grandparents or their own country of birth was known. On the next slide, these geographic locations were used to color the location of each of 1,387 people in "genetic space." Instead of latitude and longitude on a geographic map, their first two principal components were used: these components summarize the 500,000 SNPs typed for each person.

Novembre et al., 2008



Novembre et al., 2008

As a follow-up, the authors took the genetic profile of each person and used it to predict their latitude and longitude, and plotted these on a geographic map. These predicted positions are colored by the country of origin of each person.



Y SNP Data Haplogroups

Another set of SNP data, this time from around the world, is available for the Y chromosome. These data were collected for the 1000 Genomes project (http://www.1000genomes.org/): there are 26 populations:

East Asia: CDX. Chinese Dai in Xishuangbanna; CHB. Han Chinese in Beijing; JPT. Japanese in Tokyo; KHV. Kinh in Ho Chi Minh City; CHS. Southern Han Chinese.

South Asian: BEB. Bengali in Bangladesh; GIH. Gujarati Indian in Houston; ITU. India Telugi in UK; PJL. Punjabi in Lahore; STU. Sri Lankan Tamil in UK.

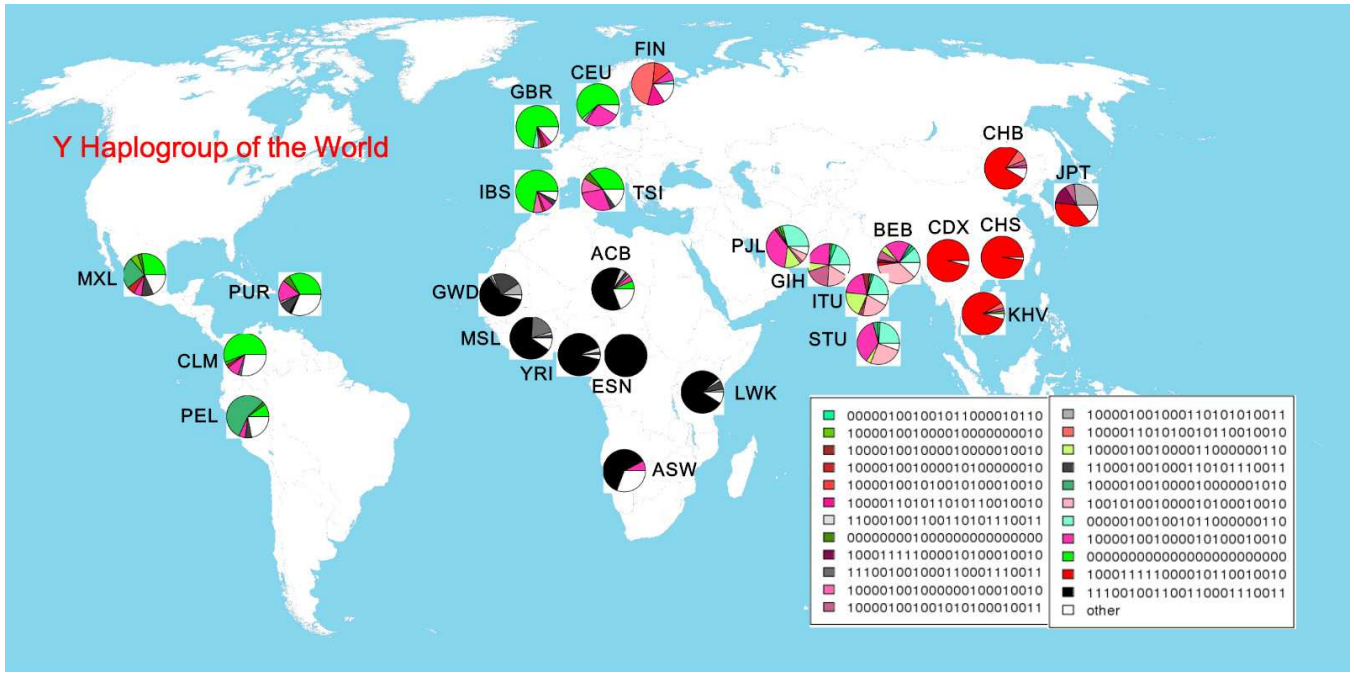
Y SNP Data Haplogroups

African: ASW. African Ancestry in Southwest US; ACB. African Caribbean in Barbados; ESN. Esan in Nigeria; GWD. Gambian in the Gambia; LWK. Luthya in Kenya; MSL. Mende in Sierra Leone; YRI. Yoruba in Nigeria.

European: GBR. British in UK; FIN. Finnish in Finland; IBS. Iberian in Spain; TSI. Toscani in Italy; CEU. Utah residents with European ancestry.

Americas: CLM. Columbian in Medellin; MXL. Mexican in Los Angeles; PEL. Peruvian in Lima, PUR. Puerto Rican in Puerto Rico.

Y SNP Data Haplogroups



Migration History of Early Humans

An interesting video of the migration of early humans is available at:

http://www.bradshawfoundation.com/journey/

Migration Map of Early Humans

https://genographic.nationalgeographic.com/human-journey/

This map summarizes the migration patterns of early humans.

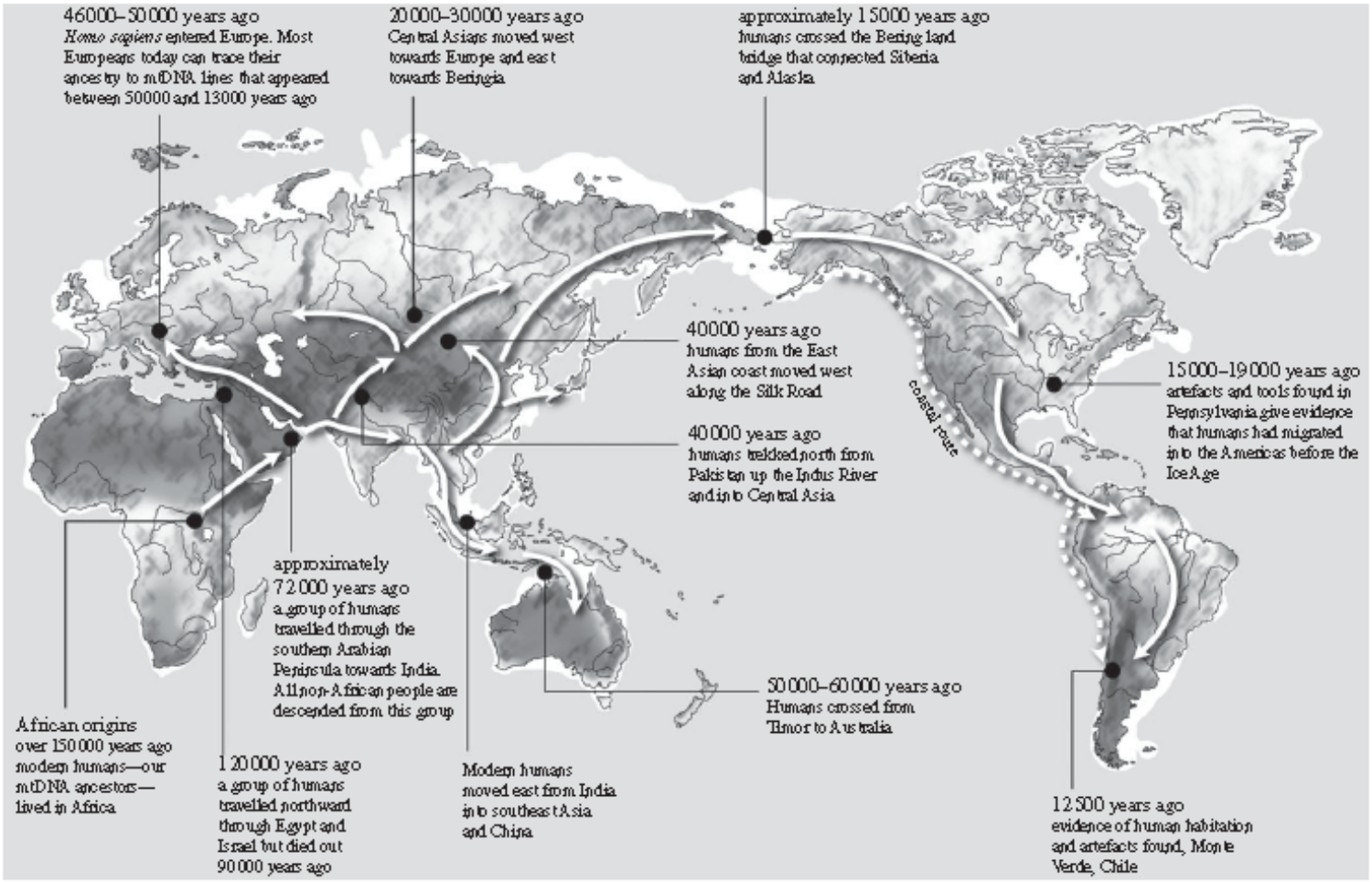
Migration Map of Early Humans

The map on the next slide, based on mitochondrial genetic profiles, is taken from:

Oppenheimer S. 2012. Out-of-Africa, the peopling of continents and islands: tracing uniparental gene trees across the map. Phil. Trans. R. Soc. B (2012) 367, 770-784 doi:10.1098/rstb.2011.0306.

The first two pages of this paper give a good overview, and they contain this quote: "The finding of a greater genetic diversity within Africa, when compared with outside, is now abundantly supported by many genetic markers; so Africa is the most likely geographic origin for a modern human dispersal."

Migration Map of Early Humans



Forensic Implications

What does the theory about the spread of modern humans tell us about how to interpret matching profiles?

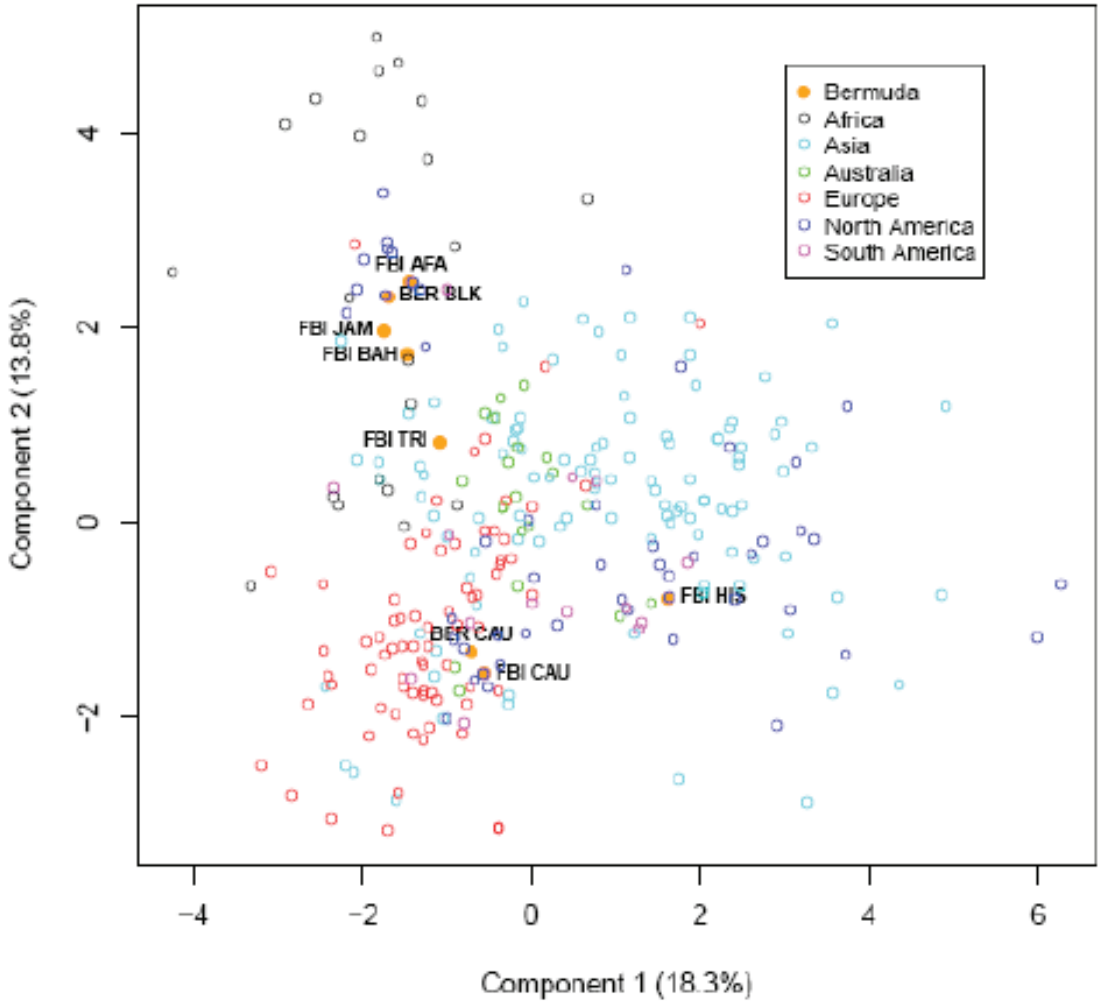
Matching probabilities should be bigger within populations, and more similar among populations that are closer together in time.

Forensic allele frequencies are consistent with the theory of human migration patterns.

Forensic STR PCA Map

A large collection of forensic STR allele frequencies was used to construct the principal component map on the next page. Also shown are some data collected by forensic agencies in the Caribbean, and by the FBI. The Bermuda police has been using FBI data - does this seem to be reasonable?

Forensic STR PCA Map



Genetic Distances

Forensic allele frequencies were collected from 21 populations. The next slides list the populations and show allele frequencies for the Gc marker. This has only three alleles, A, B, C.

The matching proportions within each population, and between each pair of populations, were calculated. These allow distances ("theta" or β) to be calculated for each pair of populations, say 1 and 2: $\hat{\beta}_{12} = (\tilde{M}_1 + \tilde{M}_2 - 2\tilde{M}_{12})/[2(1 - \tilde{M}_{12})]$.

$\tilde{M}_1$: two alleles taken randomly from population 1 are the same type.
$\tilde{M}_1$: two alleles taken randomly from population 1 are the same type.
$\tilde{M}_{12}$: an allele taken randomly from population 1 matches an allele taken randomly from population 2.

Published Gc frequencies

Symbol	Description
AFA	FBI African-American
AL1	North Slope Alaskan
AL2	Bethel-Wade Alaskan
ARB	Arabic
CAU	FBI Caucasian
CBA	Coimbran
DUT	Dutch Caucasian
GAL	Galician
HN1	Hungarian
HN2	Hungarian
IT2	Italian

Published Gc frequencies

Symbol	Description
IT4	Italian
KOR	Korean
NAV	Navajo
NBA	North Bavarian
PBL	Pueblo
SEH	FBI Southeastern Hispanic
SOU	Sioux
SPN	Spanish
SWH	FBI Southwestern Hispanic
SWI	Swiss Caucasian

Gc allele frequencies

Popn.	Sample size	A	B	C
AFA	145	.338	.237	.423
AL1	96	.177	.489	.334
AL2	112	.236	.451	.313
ARB	94	.133	.441	.425
CAU	148	.114	.456	.429
CBA	119	.159	.533	.306
DUT	155	.106	.422	.471
GAL	143	.140	.448	.413
HN1	345	.106	.457	.438
HN2	163	.097	.448	.454
IT2	374	.139	.454	.408

Gc allele frequencies

Popn.	Sample size	A	B	C
IT4	200	.302	.163	.535
KOR	116	.310	.422	.267
NAV	81	.105	.240	.654
NBA	150	.133	.383	.484
PBL	103	.102	.374	.524
SEH	94	.165	.447	.389
SOU	64	.055	.422	.524
SPN	132	.118	.474	.409
SWH	96	.156	.437	.407
SWI	100	.135	.465	.400

Distances based on Gc

	AFA	AL1	AL2	ARB	CAU	CBA	DUT	GAL	HN1	HN2
AL1	.201									
AL2	.163	.000								
ARB	.224	.002	.016							
CAU	.303	.020	.046	.008						
CBA	.309	.017	.034	.022	.009					
DUT	.341	.039	.070	.021	.000	.017				
GAL	.295	.015	.037	.007	.000	.004	.002			
HN1	.339	.040	.072	.025	.001	.013	.000	.002		
HN2	.348	.041	.073	.024	.000	.016	.000	.003	.000	
IT2	.304	.023	.048	.015	.000	.004	.002	.000	.001	.002

Distances based on Gc

	AFA	AL1	AL2	ARB	CAU	CBA	DUT	GAL	HN1	HN2
IT4	.088	.029	.022	.032	.085	.098	.111	.081	.120	.117
KOR	.074	.051	.026	.082	.139	.122	.175	.128	.179	.179
NAV	.242	.060	.080	.028	.054	.103	.063	.061	.075	.070
NBA	.278	.017	.041	.002	.000	.018	.004	.001	.007	.006
PBL	.178	.033	.044	.015	.051	.085	.067	.053	.077	.073
SEH	.254	.001	.015	.000	.002	.005	.014	.000	.014	.015
SOU	.294	.035	.062	.008	.010	.046	.012	.015	.020	.016
SPN	.315	.022	.048	.012	.000	.005	.000	.000	.000	.000
SWH	.269	.004	.022	.000	.000	.004	.008	.000	.009	.009
SWI	.298	.013	.035	.007	.000	.002	.002	.000	.002	.003

Distances based on Gc

	IT2	IT4	KOR	NAV	NBA	PBL	SEH	SOU	SPN	SWH
IT4	.098									
KOR	.145	.026								
NAV	.072	.048	.143							
NBA	.005	.067	.127	.034						
PBL	.066	.016	.088	.003	.032					
SEH	.004	.052	.089	.054	.003	.038				
SOU	.021	.067	.148	.011	.001	.021	.019			
SPN	.000	.093	.144	.066	.002	.061	.003	.016		
SWH	.001	.060	.102	.053	.000	.040	.000	.014	.000	
SWI	.000	.079	.125	.062	.001	.054	.000	.016	.000	.000

Clustering populations

Populations can be clustered on the basis of the genetic distances between them. For short-term evolution (among human populations) the simple UPGMA method performs satisfactorily. The closest pair of populations are clustered, and then distances recomputed from each other population to this cluster. Then the process continues.

Look at four of the populations:

	AFA	CAU	SEH	NAV
AFA	–			
CAU	0.303	–		
SEH	0.254	0.002	–	
NAV	0.242	0.054	0.054	–

Clustering populations

The closest pair is CAU/SEH. Cluster them, and compute distances from the other two to this cluster:

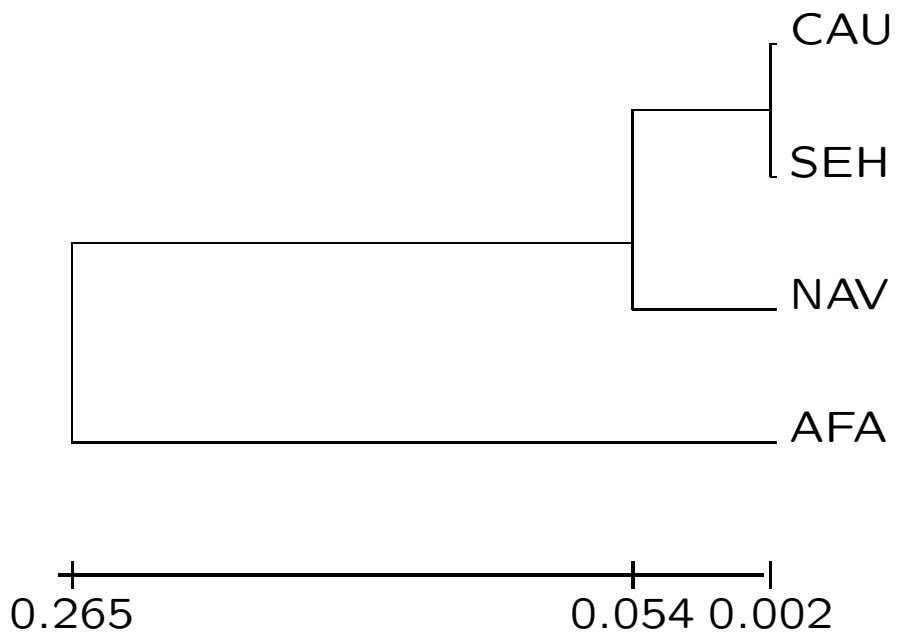
AFA distance = (0.303+0.254)/2 = 0.278
NAV distance = (0.054+0.054)/2 = 0.054

The new distance matrix is

	AFA	CAU/SEH	NAV
AFA	–		
CAU/SEH	0.278	–	
NAV	0.242	0.054	–

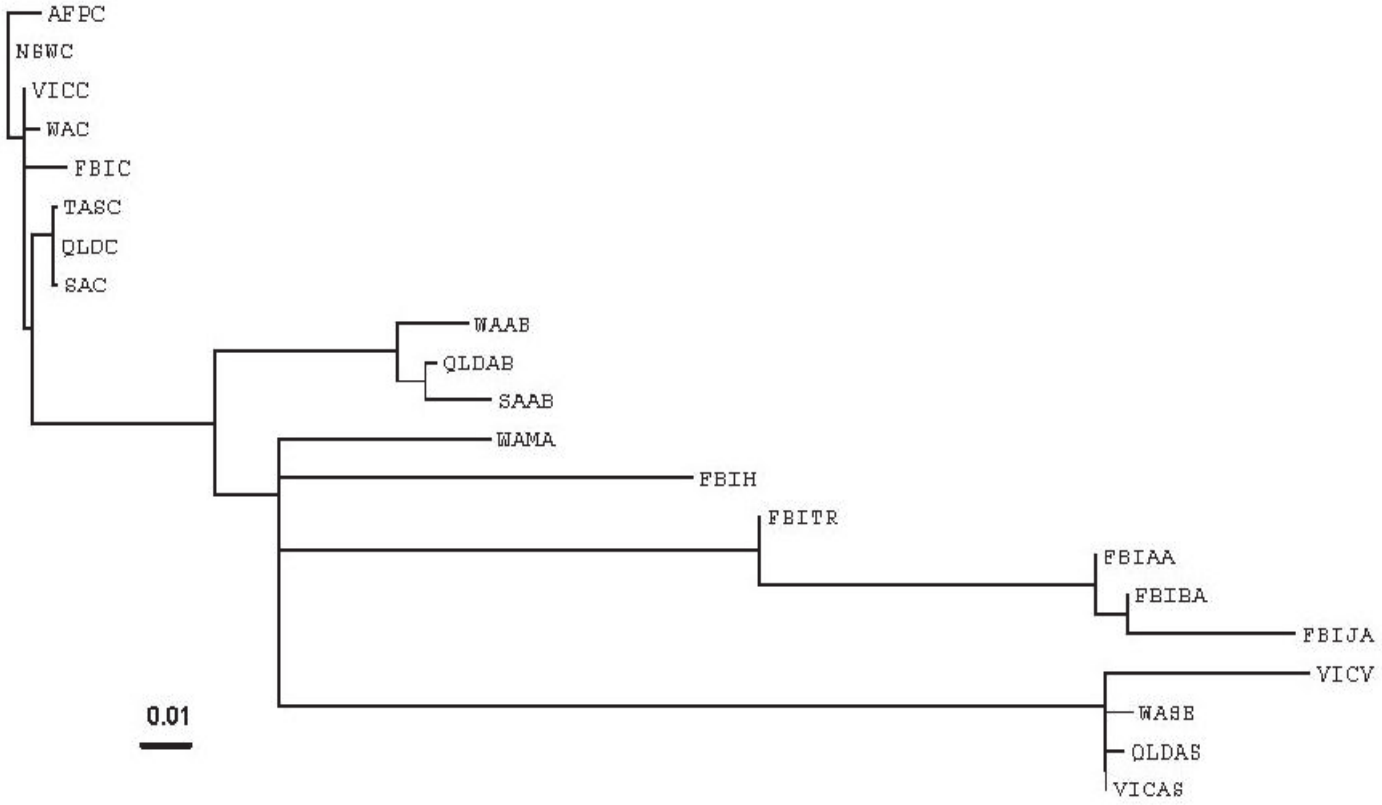
and the next shortest distance is between NAV and CAU/SEH.

Gc UPGMA Dendrogram



Australian STR Data

Australian Values



115

Worldwide Survey of STR Data

Published allele frequencies for 24 STR loci were obtained for 446 populations. For each population i, the within-population matching proportion $\tilde{M}_i$ was calculated. Also the average $\tilde{M}_B$ of all the between-population matching proportions. The "θ" for each population is calculated as $\hat{\beta}_i = (\tilde{M}_i - \tilde{M}_B)/(1 - \tilde{M}_B)$. These are shown on the next slide, ranked from smallest to largest and colored by continent.

Africa: black; America: red; South Asia: orange; East Asia: yellow; Europe: blue; Latino: turquoise; Middle East: grey; Oceania: green.

Worldwide Survey of STR Data