POPULATION STRUCTURE

and

RELATEDNESS

ALLELE MATCHING

Allele Matching

Forensic genetics is concerned with matching of genetic profiles from evidence and from persons of interest. Profile match probabilities rest on the probabilities of matching among the alleles constituting the profiles.

Allele matching can refer to alleles within an individual (inbreeding), between individuals within a population (relatedness) and between populations (population structure). In all these cases there are parameters that describe profile match probabilities, and these parameters can be estimated by comparing the observed matching for a target set of alleles with that between a comparison set.

Allele Matching Within Individuals

The inbreeding coefficient for an individual is the probability it receives two alleles at a locus, one from each parent, that are *identical by descent.*

What can be observed, however, is identity in state. An individual is either homozygous or heterozygous at a locus: the two alleles either match or miss-match at that locus. The proportion of matching alleles at a locus is either zero or one, not a very informative statistic, but the proportion of an individual's loci that are homozygous may be informative for their inbreeding status.

There is still a need for a reference: many loci will be homozygous even for non-inbred individuals. Therefore we compare the proportion of loci with matching alleles for an individual with the matching proportion for pairs of alleles taken one from each of two individuals: is allele matching higher within than between individuals?

Inbreeding

If $\tilde{M}_j$ is the observed proportion of loci with matching alleles (i.e. homozygous) for individual j, and if $\tilde{M}_S$ is the observed proportion of matching alleles, one from each of two individuals in the population, then the within-population inbreeding coefficient f_j is estimated as

$$\hat{f}_j = \frac{\tilde{M}_j - \tilde{M}_S}{1 - \tilde{M}_S}$$

Note that this can be negative for individuals with high degrees of heterozygosity.

The average of these estimates over all the individuals in a sample from a population estimates the within-population inbreeding coefficient f:

$$\hat{f} = \frac{\tilde{M}_I - \tilde{M}_S}{1 - \tilde{M}_S}$$

where $\tilde{M}_I = \sum_{j=1}^{n} \tilde{M}_j/n$. Hardy-Weinberg equilibrium corresponds to $f = 0$.

Allele Matching Between Individuals

How can we tell if a pair of individuals has a high degree of allele matching? What does "high" mean?

We assess relatedness of individuals within a population by comparing their degree of allele matching with the average degree for all pairs of individuals in that population.

Allele Matching Between Individuals

If $\tilde{M}_{jj'}$ is the observed proportion of loci with matching alleles, one from each of individuals j and j', and if $\tilde{M}_S$ is the average of all the $\tilde{M}_{jj'}$'s, then the within-population kinship coefficient $beta_{jj'}$ is estimated as

$$\hat{\psi}_{jj'} = \frac{\tilde{M}_{jj'} - \tilde{M}_S}{1 - \tilde{M}_S}$$

Note that this can be negative for pairs of individuals less related than the average pair-matching in the sample.

The average of these estimates over all pairs of individuals in a sample is zero, but this doesn't allow us to compare populations.

Allele Matching for Populations

We calibrated allele matching within individuals by comparison with matching between pairs of individuals.

We calibrate the allele matching between pairs of individuals by comparison with matching between pairs of populations. If $\tilde{M}^{ii'}$ is the observed proportion of loci with matching alleles, one from each of populations i and i', and if $\tilde{M}_B$ is the average of all the $\tilde{M}^{ii'}$'s, then the total kinship coefficient $\beta_{jj'}$ is estimated as

$$\hat{\psi}_{jj'} = \frac{\tilde{M}_{jj'} - \tilde{M}_B}{1 - \tilde{M}_B}$$

The average of these estimates over all pairs of individuals in a sample from a population is

$$\hat{\psi} = \frac{\tilde{M}_S - \tilde{M}_B}{1 - \tilde{M}_B}$$

This is the "θ" needed for the "theta correction" discussed below.

Within-population Matching

We can get some empirical matching proportions when we have a set of profiles. To simplify this initial discussion, consider the following data for the Y-STR locus DYS390 from the NIST database:

	Population				
Allele	Afr.Am.	Cauc.	Hisp.	Asian	Total
20	4	1	1	0	6
21	176	4	17	1	198
22	43	45	14	17	119
23	36	116	50	17	219
24	56	145	129	21	351
25	23	46	21	36	126
26	3	2	2	4	11
27	0	0	2	0	2
Total	341	359	236	96	1032

Within- and Between-population Matching for DYS390

Within the African-American sample there are $341 \times 340 = 115,940$ pairs of profiles and the number of between individual-pair matches is

$$4 \times 3 + 176 \times 175 + 43 \times 42 + 36 \times 35 + 56 \times 55 + 23 \times 22 + 3 \times 2 = 37,470$$

so the within-population matching proportion is $37,470/115,940 = 0.323$.

Between the African-American and Caucasian samples, there are $341 \times 359 = 122,419$ pairs of profiles and the number of matches is

$$4 \times 1 + 176 \times 4 + 43 \times 45 + 36 \times 116 + 56 \times 145 + 23 \times 4 + 3 \times 2 = 12,403$$

so the between-population matching proportion is $12,403/122,419 = 0.101$.

Two-locus counts in NIST African-American Data for DYS390, DYS391

DYS390	DYS391	Count n_g	$n_g(n_g-1)$
22	10	34	1122
22	11	9	72
24	10	15	210
24	11	39	1482
24	12	1	0
24	9	1	0
23	10	19	342
23	11	14	182
23	12	3	6
21	10	157	24492
21	11	15	210
21	12	2	2
21	9	1	0
21	13	1	0
25	10	11	110
25	11	12	132
26	10	1	0
26	11	2	2
20	10	1	0
20	11	2	2
20	12	1	0

Two-locus Matches

The within-population matching proportion for the African-American sample is 28,366/115,940=0.245.

The within-population matching proportion for the Caucasian sample is 18,536/128,522=0.144.

The between-population matching proportion for the African-American and Caucasian samples is 8,347/122,419=0.068.

There is a clear decrease in matching between populations from within populations.

Will match probabilities keep decreasing?

How do these match probabilities address the observation of Donnelly:

> "after the observation of matches at some loci, it is relatively much more likely that the individuals involved are related (precisely because matches between unrelated individuals are unusual) in which case matches observed at subsequent loci will be less surprising. That is, knowledge of matches at some loci will increase the chances of matches at subsequent loci, in contrast to the independence assumption."

Donnelly P. 1995. Heredity 75:26-64.

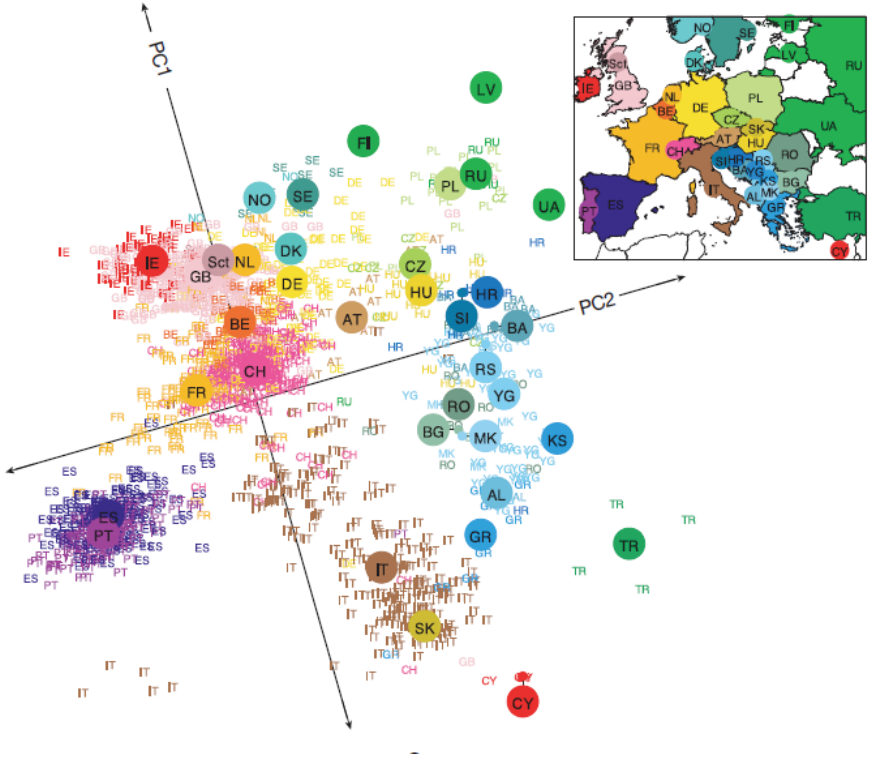
POPULATION STRUCTURE

Human Populations: History and Structure

"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." 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

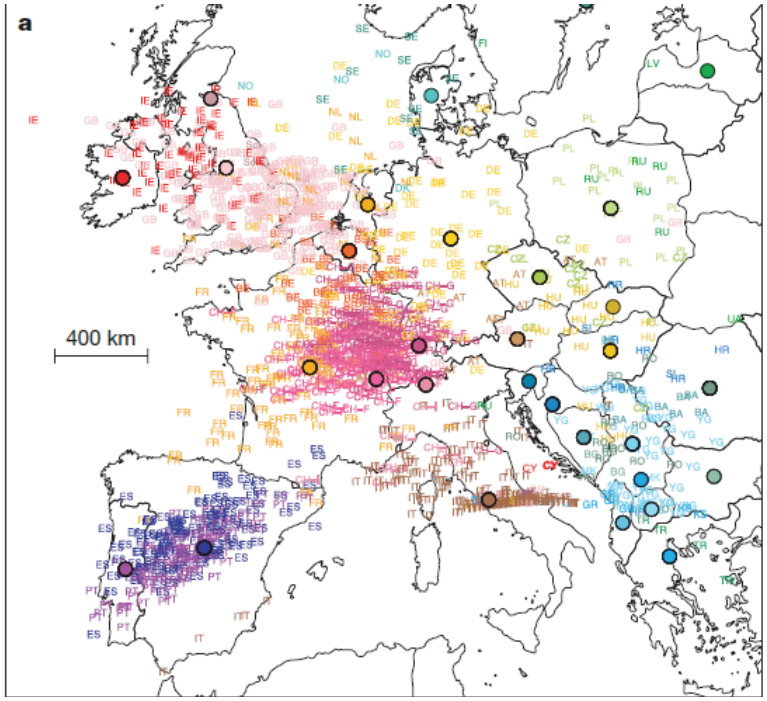
The authors collected "SNP" (single nucleotide polymorphism) data on over people living in Europe. Either the country of origin of the people's grandparents or their own country of birth was known. On the next slide, these geographic locations were used to color the location of each of 1,387 people in "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.

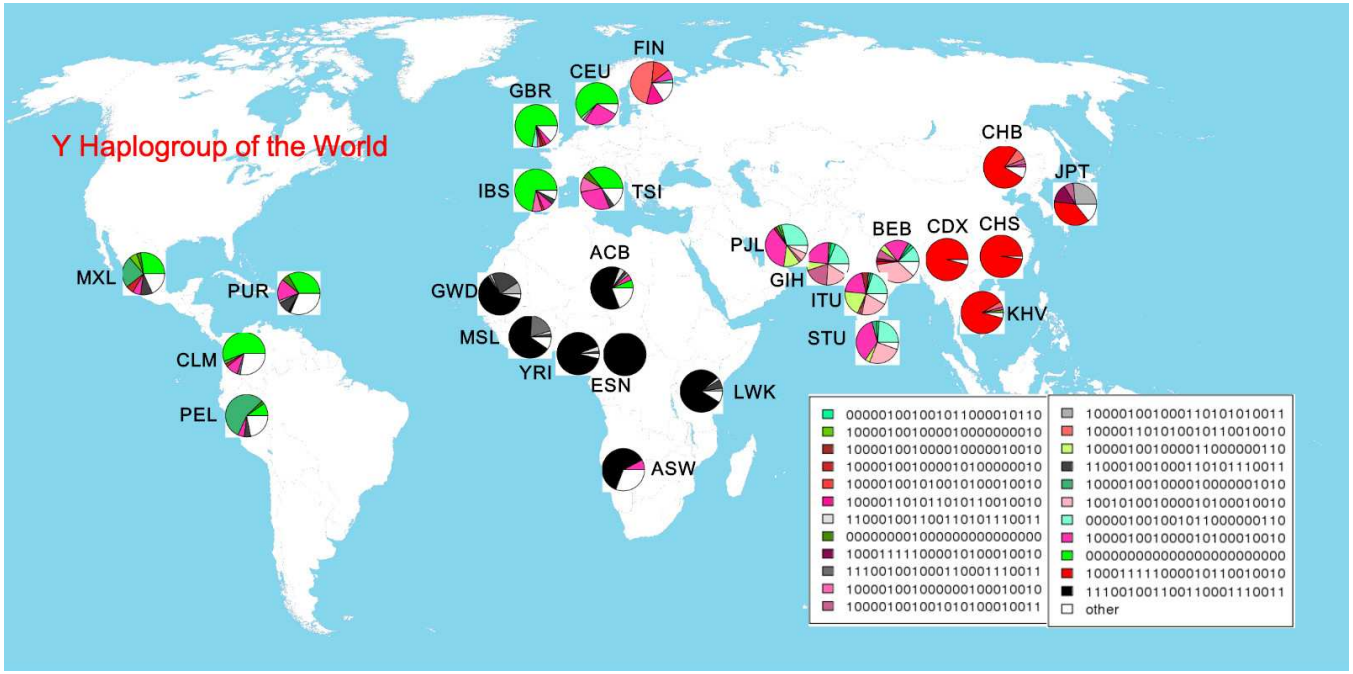


Y SNP Data Haplogroups

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 (5), South Asian (5), African (7), European (5), Americas (4).

Y SNP Data Haplogroups



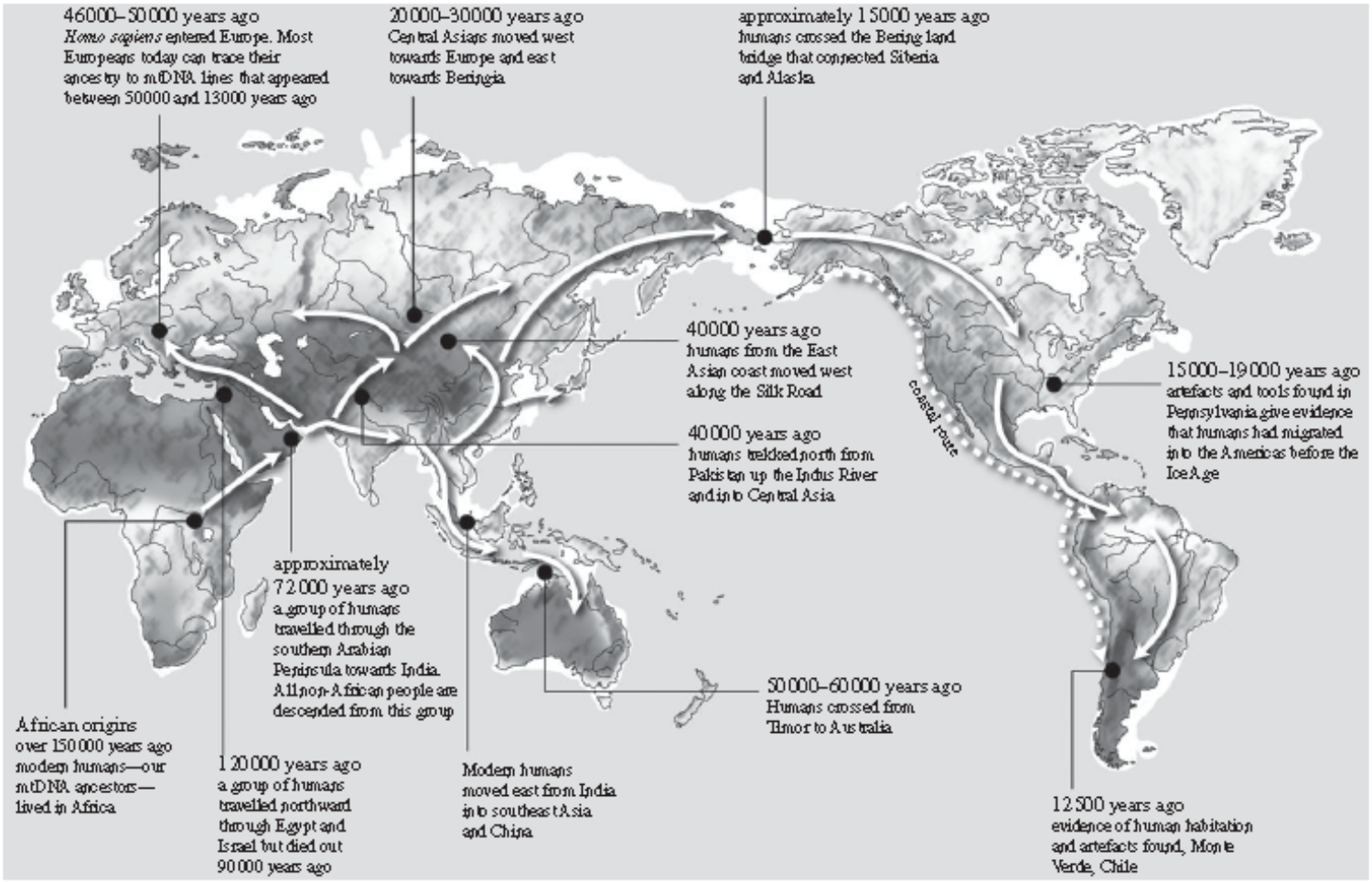
Migration Map of Early Humans

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?

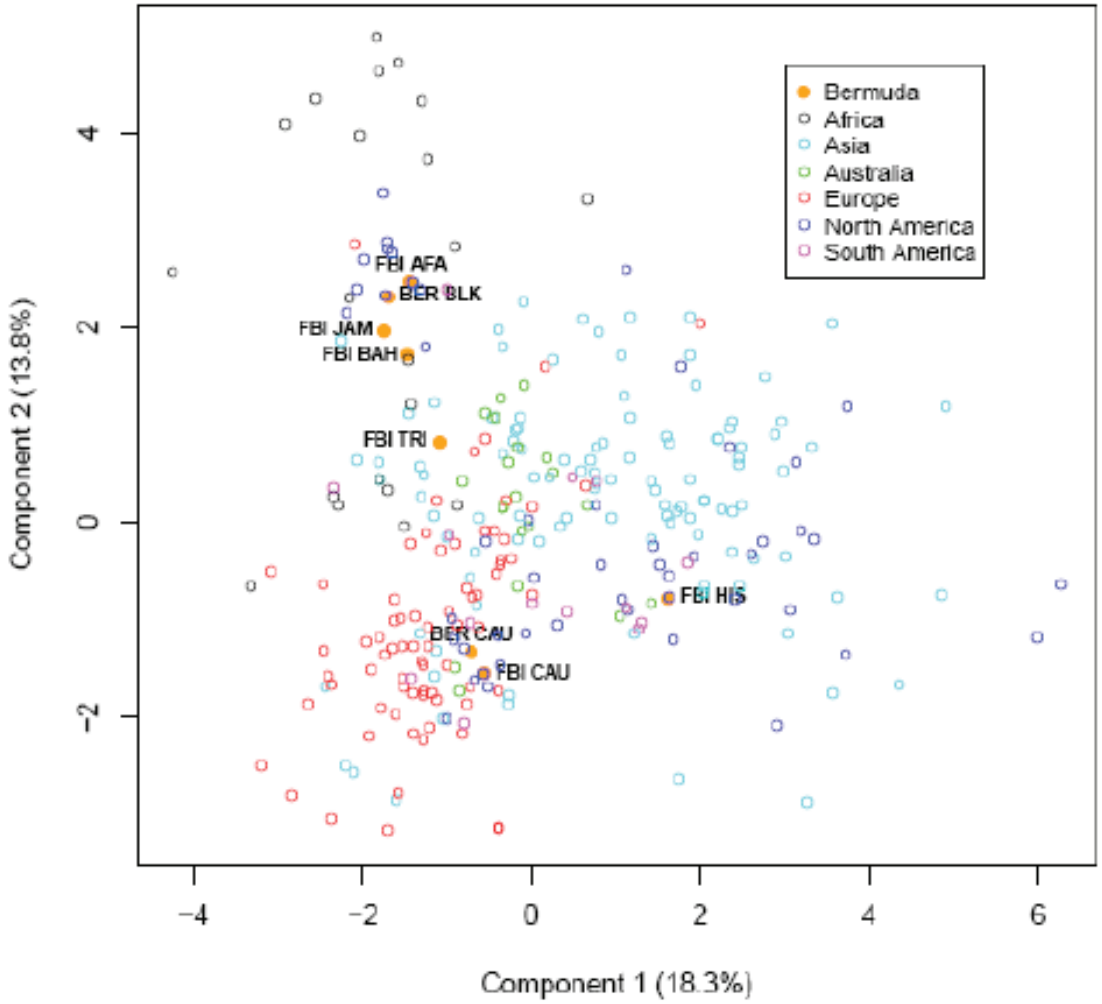
Matching probabilities should be bigger within populations, and more similar among populations that are closer together in time.

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 ψ) to be calculated for each pair of populations, say 1 and 2: $\hat{\psi}_{12} = ([\tilde{M}_1 + \tilde{M}_2]/2 - \tilde{M}_{12})/(1 - \tilde{M}_{12})$.

$\tilde{M}_1$: 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	Symbol	Description
AFA	FBI African-American	IT4	Italian
AL1	North Slope Alaskan	KOR	Korean
AL2	Bethel-Wade Alaskan	NAV	Navajo
ARB	Arabic	NBA	North Bavarian
CAU	FBI Caucasian	PBL	Pueblo
CBA	Coimbran	SEH	FBI Southeastern Hispanic
DUT	Dutch Caucasian	SOU	Sioux
GAL	Galician	SPN	Spanish
HN1	Hungarian	SWH	FBI Southwestern Hispanic
HN2	Hungarian	SWI	Swiss Caucasian
IT2	Italian		

Gc allele frequencies

Popn.	Sample size	A	B	C	Popn.	Sample size	A	B	C
AFA	145	.338	.237	.423	IT4	200	.302	.163	.535
AL1	96	.177	.489	.334	KOR	116	.310	.422	.267
AL2	112	.236	.451	.313	NAV	81	.105	.240	.654
ARB	94	.133	.441	.425	NBA	150	.133	.383	.484
CAU	148	.114	.456	.429	PBL	103	.102	.374	.524
CBA	119	.159	.533	.306	SEH	94	.165	.447	.389
DUT	155	.106	.422	.471	SOU	64	.055	.422	.524
GAL	143	.140	.448	.413	SPN	132	.118	.474	.409
HN1	345	.106	.457	.438	SWH	96	.156	.437	.407
HN2	163	.097	.448	.454	SWI	100	.135	.465	.400
IT2	374	.139	.454	.408					

Clustering populations

Populations can be clustered on the basis of the genetic distances β_{ij} between each pair i, j. 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:

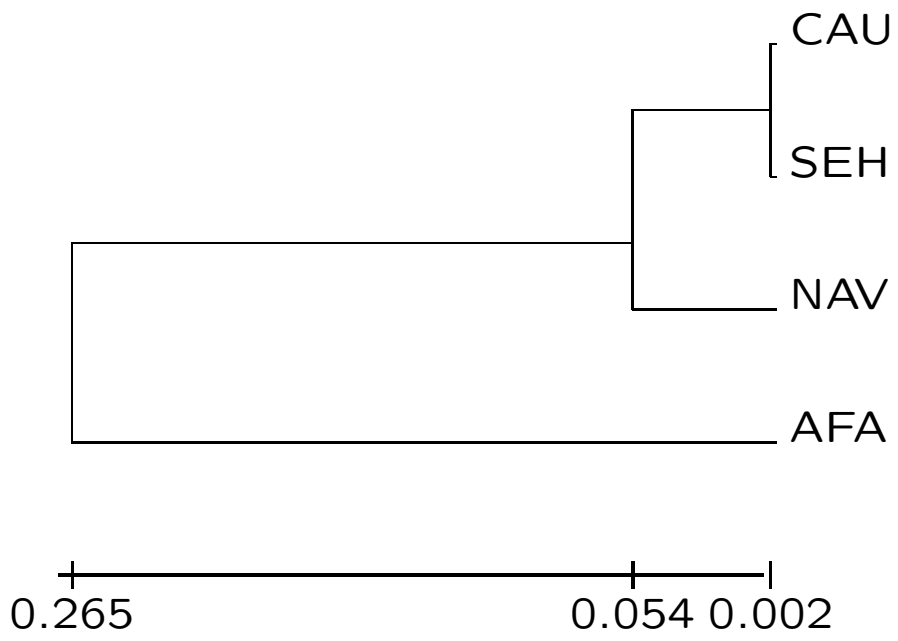
AFA distance = (0.303+0.254)/2 = 0.278
NAV distance = (0.054+0.054)/2 = 0.054

The new distance matrix is

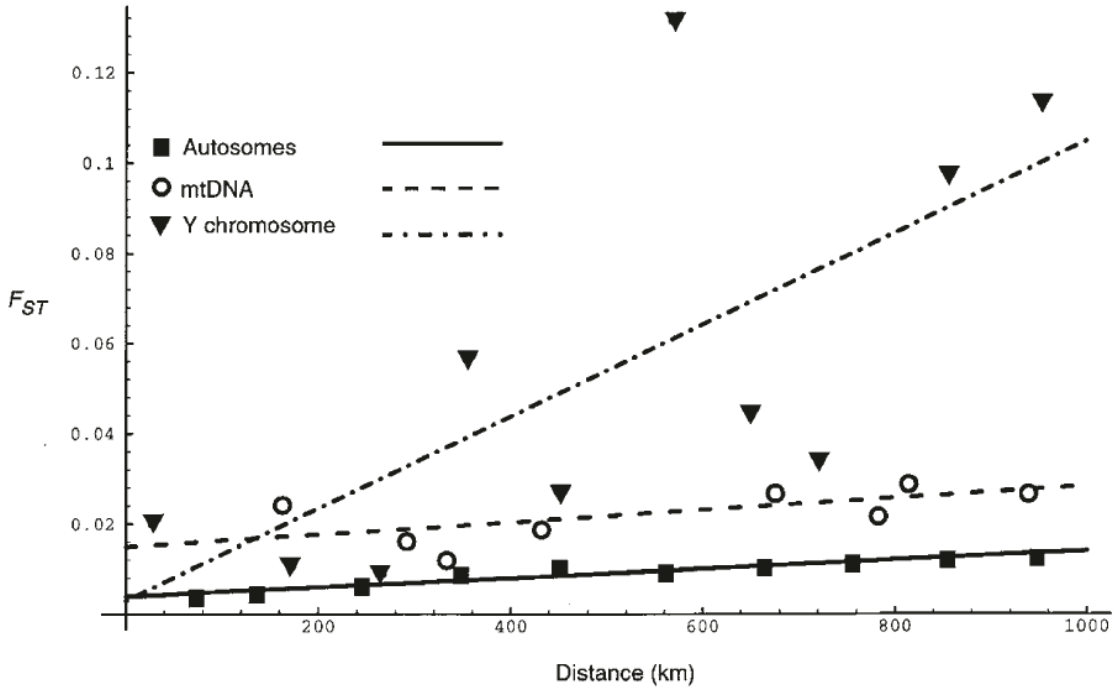
	AFA	CAU/SEH	NAV
AFA	–		
CAU/SEH	0.278	–	
NAV	0.242	0.054	–

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

Gc UPGMA Dendrogram



Human Migration Rates



Suggests higher migration rate for human females among 14 African populations.

Seielstad MT, Minch E, Cavalli-Sforza LL. 1998. Nature Genetics 20:278-280.

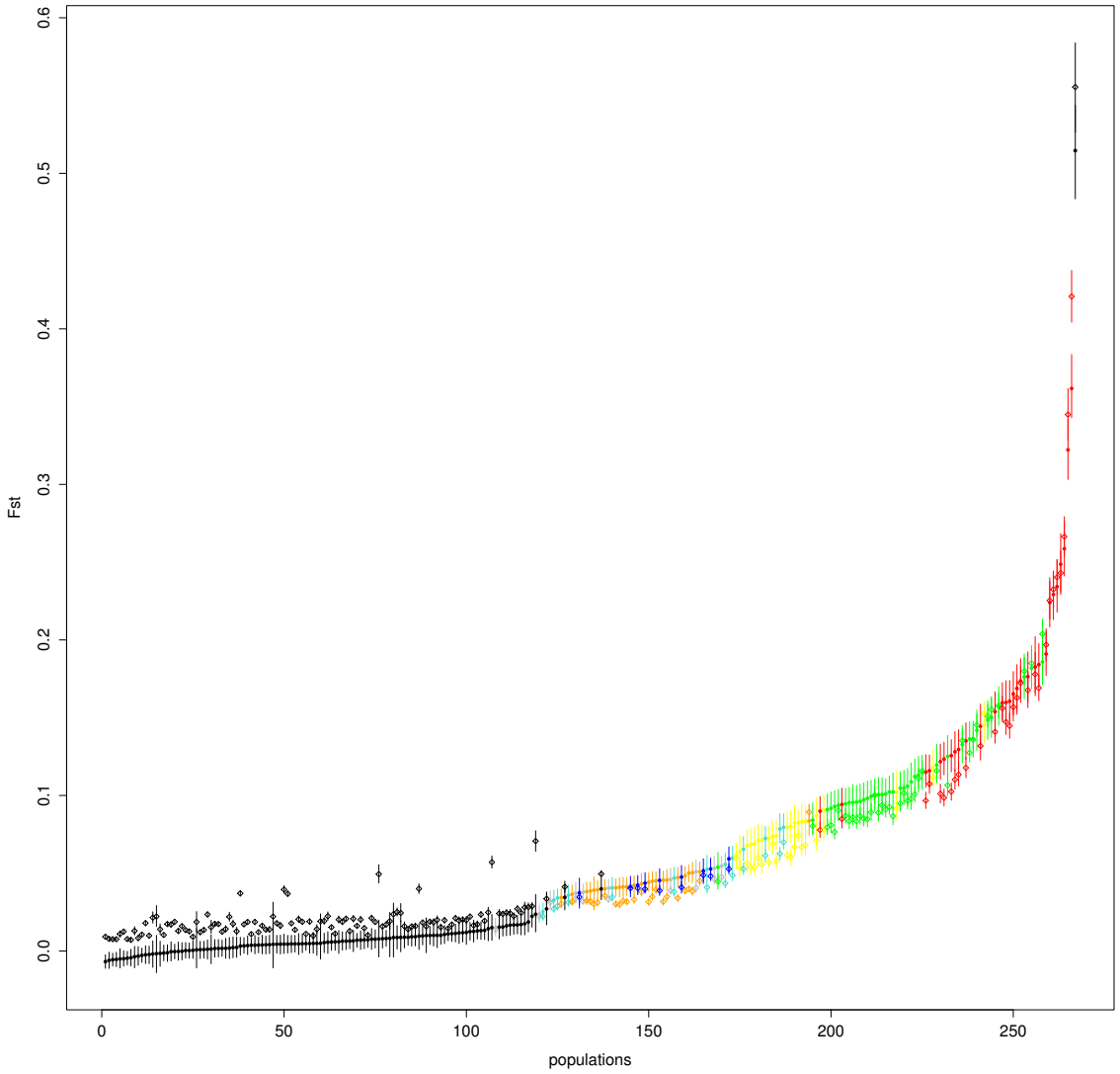
Worldwide Survey of STR Data

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 $\hat{\psi}_i = (\tilde{M}_i - \tilde{M}_B)/(1 - \tilde{M}_B)$. 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.

Buckleton JS, Curran JM, Goudet J, Taylor D, Thiery A, Weir BS. 2016. Forensic Science International: Genetics 23:91-100.

Worldwide Survey of STR Data



Match Probabilities

The β estimates for population structure provide numerical values to substitute for θ into the Balding-Nichols match probabilities when database sample allele frequencies are used for the population values p_A.

For AA homozygotes:

$$\Pr(AA|AA) = \frac{[3\theta + (1-\theta)p_A][2\theta + (1-\theta)p_A]}{(1+\theta)(1+2\theta)}$$

and for AB heterozygotes

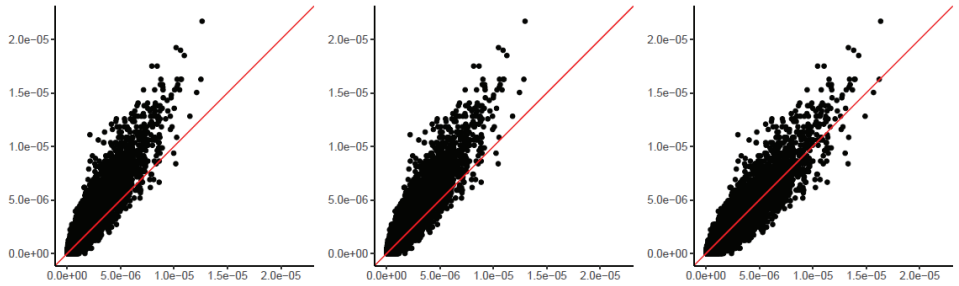
$$\Pr(AB|AB) = \frac{2[\theta + (1-\theta)p_A][\theta + (1-\theta)p_B]}{(1+\theta)(1+2\theta)}$$

These match probabilities are greater than the profile probabilities $\Pr(AA), \Pr(AB)$.

Balding DJ, Nichols RA. 1994. Forensic Science International 64:125-140.

Multi-locus Match Probabilities

In this figure we compare the observed 5-locus match proportions with the products of five θ-corrected single-locus proportions (for three different θ values, to confirm the expectation that these single-locus "corrections" compensates for multi-locus dependencies:



$\theta = 0$ $\theta = 0.001$ $\theta = 0.01$

Five-locus match proportions (Y-axis) vs products of θ-corrected proportions (X-axis).

Balding Sampling Formula

The match probabilities on the previous slide follow from a "sampling formula": the probability of seeing an A allele if the previous n alleles have n_A of type A is

$$\Pr(A|n_A \text{ of } n) = \frac{n_A\theta + (1-\theta)p_A}{1+(n-1)\theta}$$

For example, some conditional probabilitites are

$$\Pr(A) = p_A$$

$$\Pr(A|A) = [\theta + (1-\theta)p_A]$$

$$\Pr(A|AA) = \frac{[2\theta + (1-\theta)p_A]}{1+\theta}$$

$$\Pr(A|AAA) = \frac{[3\theta + (1-\theta)p_A]}{1+2\theta}$$

Balding Sampling Formula

And here are some joint probabilitites are

$$\begin{aligned}
\Pr(A) &= p_A \\
\Pr(AA) = \Pr(A)\Pr(A|A &= p_A[\theta + (1-\theta)p_A] \\
\Pr(AAA) = \Pr(AA)\Pr(A|AA) &= \frac{p_A[\theta + (1-\theta)p_A][2\theta + (1-\theta)p_A]}{1+\theta} \\
\Pr(AAAA) = \Pr(A)\Pr(A|AAA) &= \frac{p_A[\theta + (1-\theta)p_A][2\theta + (1-\theta)p_A][3\theta + (1-\theta)p_A]}{(1+2\theta)(1+\theta)}
\end{aligned}$$

Partial Matching

For autosomal markers, two profiles may:

Match:	AA, AA or AB, AB
Partially Match:	AA, AB or AB, AC
Mismatch:	AA, BB or AA, BC or AB, CD

How likely are each of these?

Database Matching

If every profile in a database is compared to every other profile, each pair can be characterized as matching, partially matching or mismatching without regard to the particular alleles. We find the probabilities of these events by adding over all allele types.

The probability P_2 that two profiles match (at two alleles) is

$$\begin{aligned} P_2 &= \sum_A \Pr(AA, AA) + \sum_{A \neq B} \Pr(AB, AB) \\ &= \frac{\sum_A p_A[\theta + (1-\theta)p_A][2\theta + (1-\theta)p_A][3\theta + (1-\theta)p_A]}{(1+\theta)(1+2\theta)} \\ &\quad + \frac{2\sum_{A \neq B}[\theta + (1-\theta)p_A][\theta + (1-\theta)p_B]}{(1+\theta)(1+2\theta)} \end{aligned}$$

Database Matching

This approach leads to probabilities P_2, P_1, P_0 of matching at 2,1,0 alleles:

$$\begin{aligned} P_2 &= \frac{1}{D}[6\theta^3 + \theta^2(1-\theta)(2+9S_2) + 2\theta(1-\theta)^2(2S_2+S_3) \\ &\quad + (1-\theta)^3(2S_2^2 - S_4)] \end{aligned}$$

$$\begin{aligned} P_1 &= \frac{1}{D}[8\theta^2(1-\theta)(1-S_2) + 4\theta(1-\theta)^2(1-S_3) \\ &\quad + 4(1-\theta)^3(S_2 - S_3 - S_2^2 + S_4)] \end{aligned}$$

$$\begin{aligned} P_0 &= \frac{1}{D}[\theta^2(1-\theta)(1-S_2) + 2\theta(1-\theta)^2(1-2S_2+S_3) \\ &\quad + (1-\theta)^3(1-4S_2+4S_3+2S_2^2-3S_4)] \end{aligned}$$

where $D = (1+\theta)(1+2\theta)$, $S_2 = \sum_A p_A^2$, $S_3 = \sum_A p_A^3$, $S_4 = \sum_A p_A^4$. For any value of θ we can predict the matching, partially matching and mismatching proportions in a database.

FBI Caucasian Matching Counts

One-locus matches in FBI Caucasian data (18,721 pairs of 13-locus profiles).

		θ				
Locus	Observed	.000	.001	.005	.010	.030
D3S1358	.077	.075	.075	.077	.079	.089
vWA	.063	.062	.063	.065	.067	.077
FGA	.036	.036	.036	.038	.040	.048
D8S1179	.063	.067	.068	.070	.072	.083
D21S11	.036	.038	.038	.040	.042	.051
D18S51	.027	.028	.029	.030	.032	.040
D5S818	.163	.158	.159	.161	.164	.175
D13S317	.076	.085	.085	.088	.090	.101
D7S820	.062	.065	.066	.068	.070	.080
CSF1PO	.122	.118	.119	.121	.123	.134
TPOX	.206	.195	.195	.198	.202	.216
THO1	.074	.081	.082	.084	.086	.096
D16S539	.086	.089	.089	.091	.094	.105

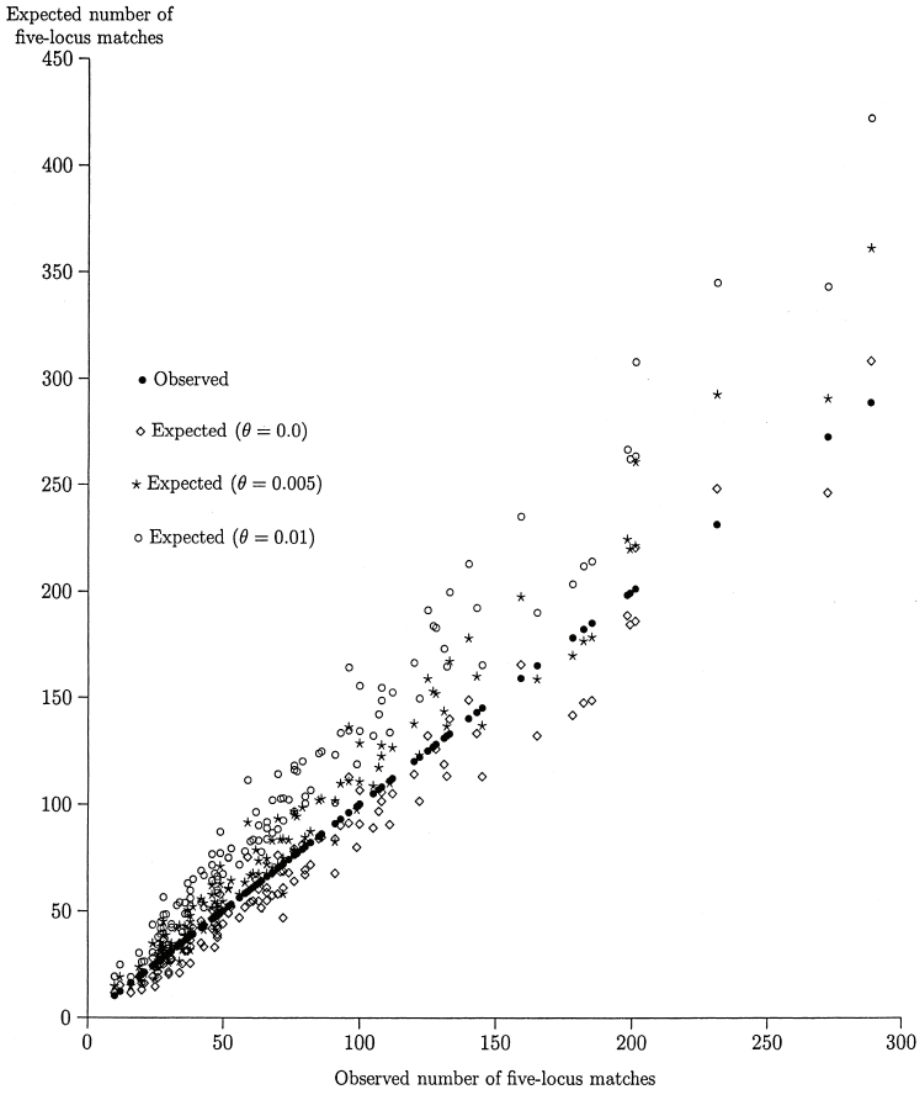
FBI Database Matching Counts

Match -ing	θ	Number of Partially Matching Loci												
		0	1	2	3	4	5	6	7	8	9	10	11	12,13
0	Obs.	0	3	18	92	249	624	1077	1363	1116	849	379	112	25, 4
	.000	0	2	19	90	293	672	1129	1403	1290	868	415	134	26, 2
	.010	0	2	14	70	236	566	992	1289	1241	875	439	148	30, 3
1	Obs.	0	12	48	203	574	1133	1516	1596	1206	602	193	43	3,
	.000	0	7	50	212	600	1192	1704	1768	1320	692	242	51	5,
	.010	0	5	40	178	527	1094	1637	1779	1393	767	282	62	6,
2	Obs.	0	7	61	203	539	836	942	807	471	187	35	2	
	.000	1	9	56	210	514	871	1040	877	511	196	45	5	
	.010	1	8	50	193	494	875	1096	969	593	239	57	6	
3	Obs.	0	6	33	124	215	320	259	196	92	16	1		
	.000	1	7	36	116	243	344	334	220	94	23	3		
	.010	0	6	35	117	256	380	387	268	120	32	4		
4	Obs.	1	5	17	29	54	82	67	16	6	0			
	.000	0	3	15	40	70	81	61	29	8	1			
	.010	0	3	15	44	81	98	78	40	12	1			
5	Obs.	0	1	2	6	12	14	6	5	0				
	.000	0	1	4	9	13	11	6	2	0				
	.010	0	1	4	11	16	15	9	3	0				
6	Obs.	0	1	0	2	2	0	0	0					
	.000	0	0	1	1	1	1	0	0					
	.010	0	0	1	2	2	1	1	0					

Predicted Matches when $n = 65,493$

Matching loci	Number of partially matching loci							
	0	1	2	3	4	5	6	7
6	4,059	37,707	148,751	322,963	416,733	319,532	134,784	24,125
7	980	7,659	24,714	42,129	40,005	20,061	4,150	
8	171	1,091	2,764	3,467	2,153	530		
9	21	106	198	163	50			
10	2	7	8	3				
11	0	0	0					
12	0	0						
13	0							

Multi-locus Matches



STR Survey: $\widehat{\psi}$ Values for Groups and Loci

	Geographic Region								
Locus	Africa	AusAb	Asian	Cauc	Hisp	IndPK	NatAm	Poly	Aver.
CSF1PO	0.003	0.002	0.008	0.008	0.002	0.007	0.055	0.026	0.011
D1S1656	0.000	0.000	0.000	0.002	0.003	0.000	0.000	0.000	0.011
D2S441	0.000	0.000	0.002	0.003	0.021	0.000	0.000	0.000	0.020
D2S1338	0.009	0.004	0.011	0.017	0.013	0.003	0.023	0.005	0.031
D3S1358	0.004	0.010	0.009	0.006	0.012	0.040	0.079	0.001	0.025
D5S818	0.002	0.013	0.009	0.008	0.014	0.018	0.044	0.007	0.029
D6S1043	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.016
D7S820	0.004	0.021	0.010	0.007	0.007	0.046	0.030	0.005	0.026
D8S1179	0.003	0.007	0.012	0.006	0.002	0.031	0.020	0.008	0.019
D10S1248	0.000	0.000	0.000	0.002	0.004	0.000	0.000	0.000	0.007
D12S391	0.000	0.000	0.000	0.003	0.020	0.000	0.000	0.000	0.010
D13S317	0.015	0.016	0.013	0.008	0.014	0.025	0.050	0.014	0.038
D16S539	0.007	0.002	0.015	0.006	0.009	0.005	0.048	0.004	0.021
D18S51	0.011	0.012	0.014	0.006	0.004	0.010	0.033	0.003	0.018
D19S433	0.009	0.001	0.009	0.010	0.014	0.000	0.022	0.014	0.023
D21S11	0.014	0.012	0.013	0.007	0.006	0.023	0.067	0.018	0.021
D22S1045	0.000	0.000	0.007	0.001	0.000	0.000	0.000	0.000	0.015
FGA	0.002	0.009	0.012	0.004	0.007	0.016	0.021	0.006	0.013
PENTAD	0.008	0.000	0.012	0.012	0.002	0.017	0.000	0.000	0.022
PENTAE	0.002	0.000	0.017	0.006	0.003	0.012	0.000	0.000	0.020
SE33	0.000	0.000	0.012	0.001	0.000	0.000	0.000	0.000	0.004
TH01	0.022	0.001	0.022	0.016	0.018	0.014	0.071	0.017	0.071
TPOX	0.019	0.087	0.016	0.011	0.007	0.018	0.064	0.031	0.035
VWA	0.009	0.007	0.017	0.007	0.012	0.022	0.028	0.005	0.023
All Loci	0.006	0.014	0.010	0.007	0.008	0.018	0.043	0.011	0.022

Buckleton JS, Curran JM, Goudet J, Taylor D, Thiery A, Weir BS. 2016. Forensic Science International: Genetics 23:91-100.

World Survey Parallel Coordinates

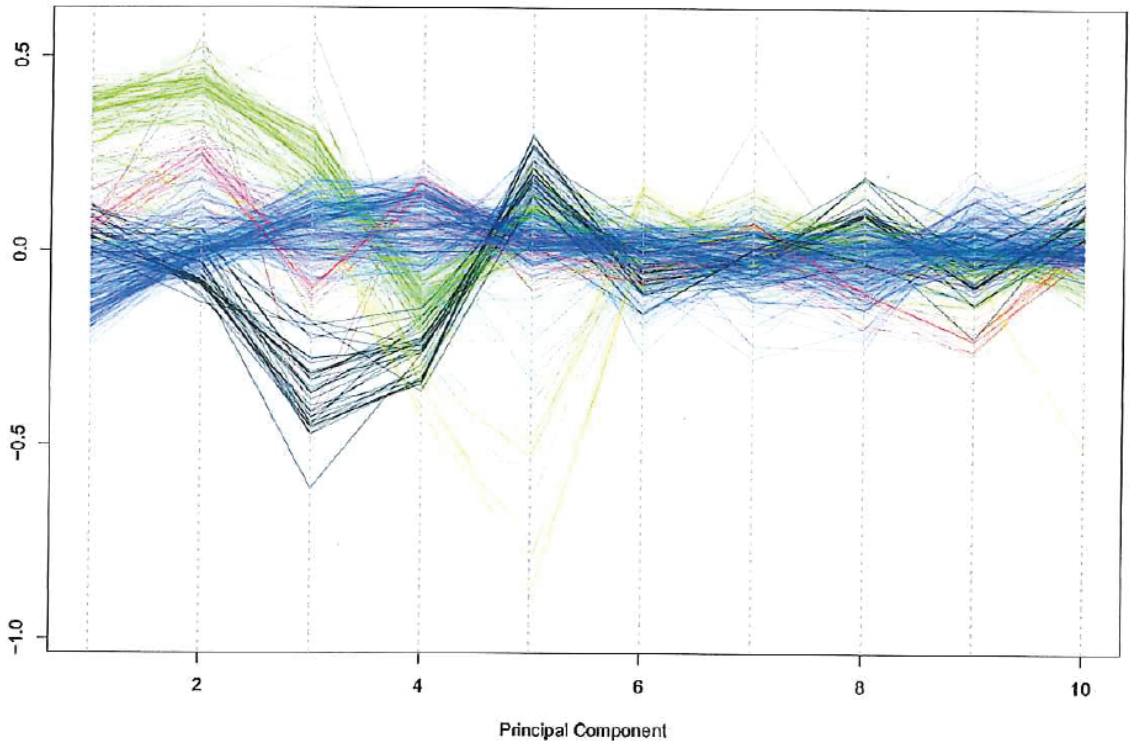


Figure 1 Parallel coordinate plot for first 10 principal coordinatess for all populations with sample sizes at least 50. Each line in the plot represents one population. Color code: Black=African, Grey=AusAb, Yellow=Asian, Blue=Caucn, Purple=Hisp, Brown=IndPk, Red=NatAm, Orange=Inuit, Brown=Andam, Green=Polyn.