Section 5:
Population Structure and Relatedness

Human Populations: History and Structure

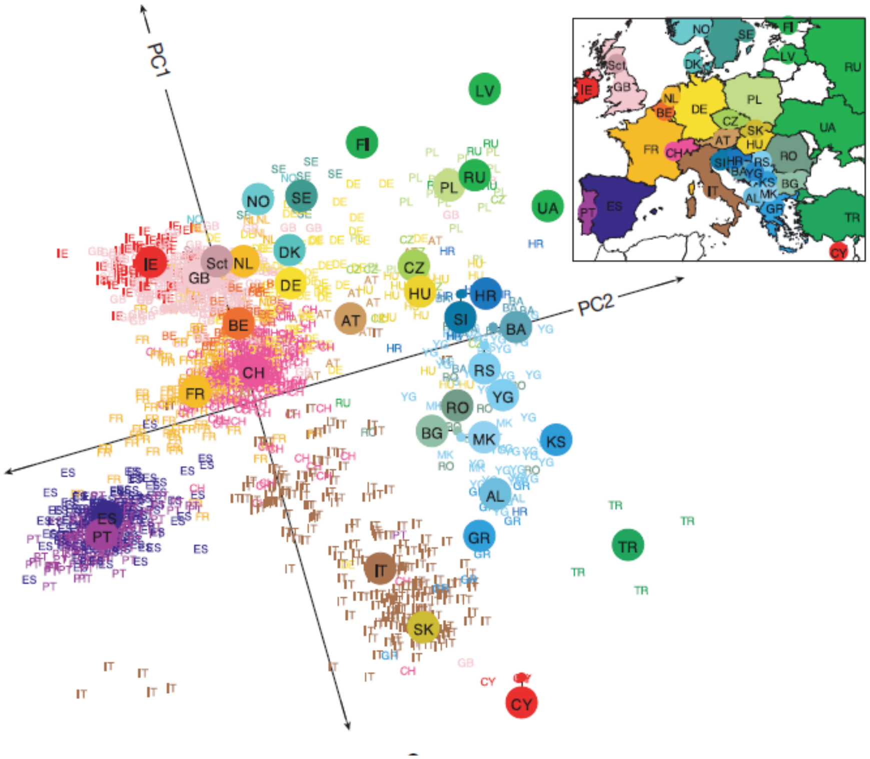
In the paper

> 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.

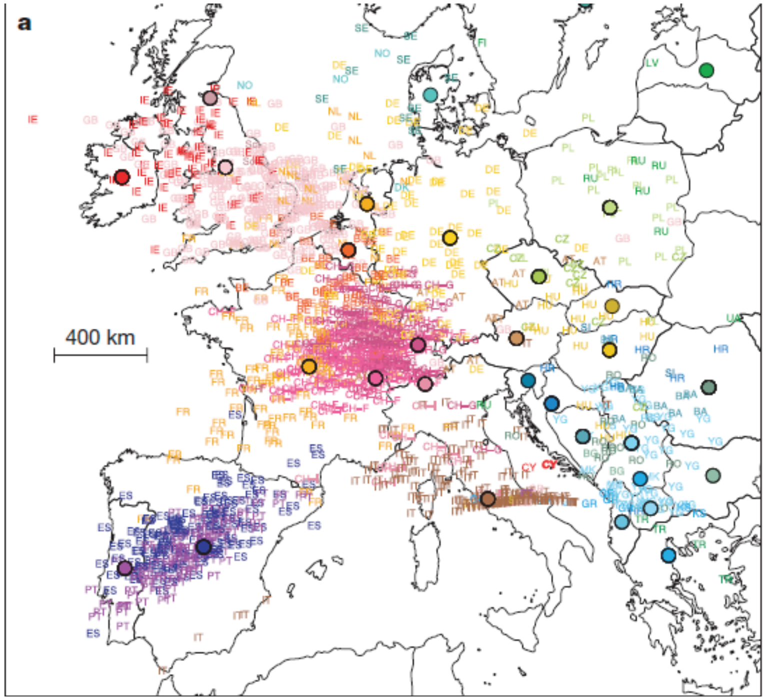
The authors collected “SNP” (single nucleotide polymorphism) data on over people living in Europe. Either the country of origin of the people’s grandparents or their own country of birth was known. On the next slide, these geographic locations were used to color the location of each of 1,387 people in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

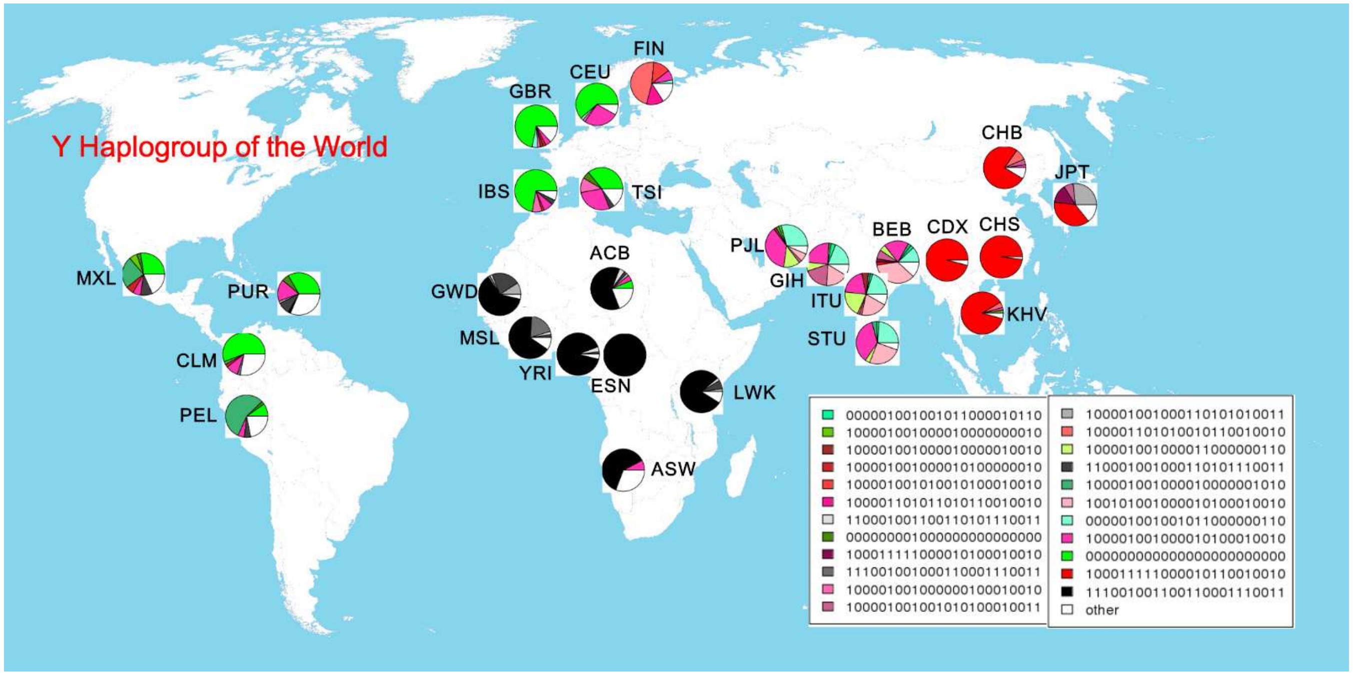


Y SNP Data Haplogroups

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

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.

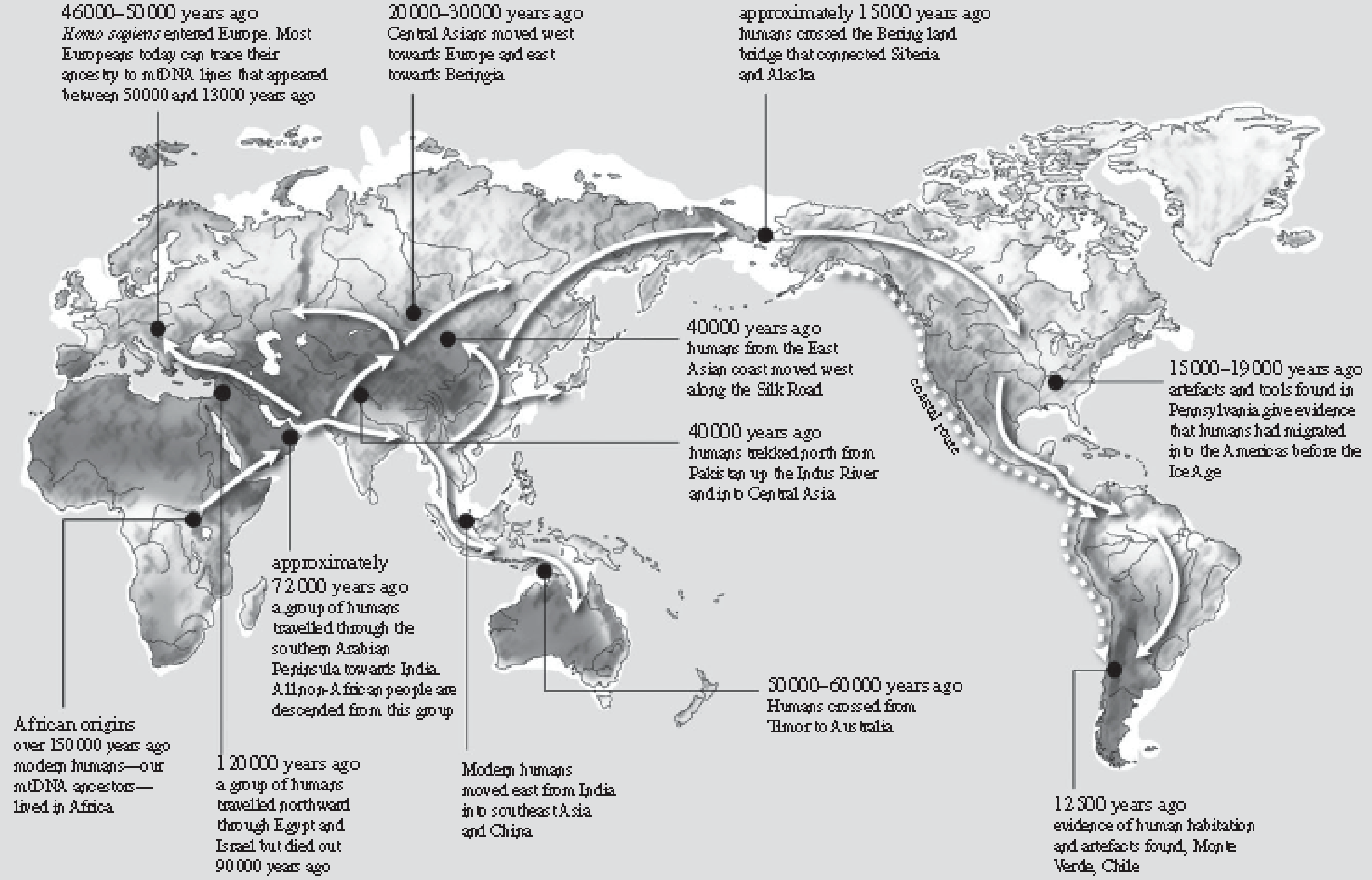
Migration Map of Early Humans

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



46000–50000 years ago
Homo sapiens entered Europe. Most Europeans today can trace their ancestry to mtDNA lines that appeared between 50000 and 13000 years ago

20000–30000 years ago
Central Asians moved west towards Europe and east towards Beringia

approximately 15000 years ago
humans crossed the Bering land bridge that connected Siberia and Alaska

40000 years ago
humans from the East Asian coast moved west along the Silk Road

40000 years ago
humans trekked north from Pakistan up the Indus River and into Central Asia

15000–19000 years ago
artefacts and tools found in Pennsylvania give evidence that humans had migrated into the Americas before the Ice Age

approximately 72000 years ago
a group of humans travelled through the southern Arabian Peninsula towards India. All non-African people are descended from this group

50000–60000 years ago
Humans crossed from Timor to Australia

African origins
over 150000 years ago modern humans—our mtDNA ancestors—lived in Africa

120000 years ago
a group of humans travelled northward through Egypt and Israel but died out 90000 years ago

Modern humans moved east from India into southeast Asia and China

12500 years ago
evidence of human habitation and artefacts found, Monte Verde, Chile

Forensic Implications

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

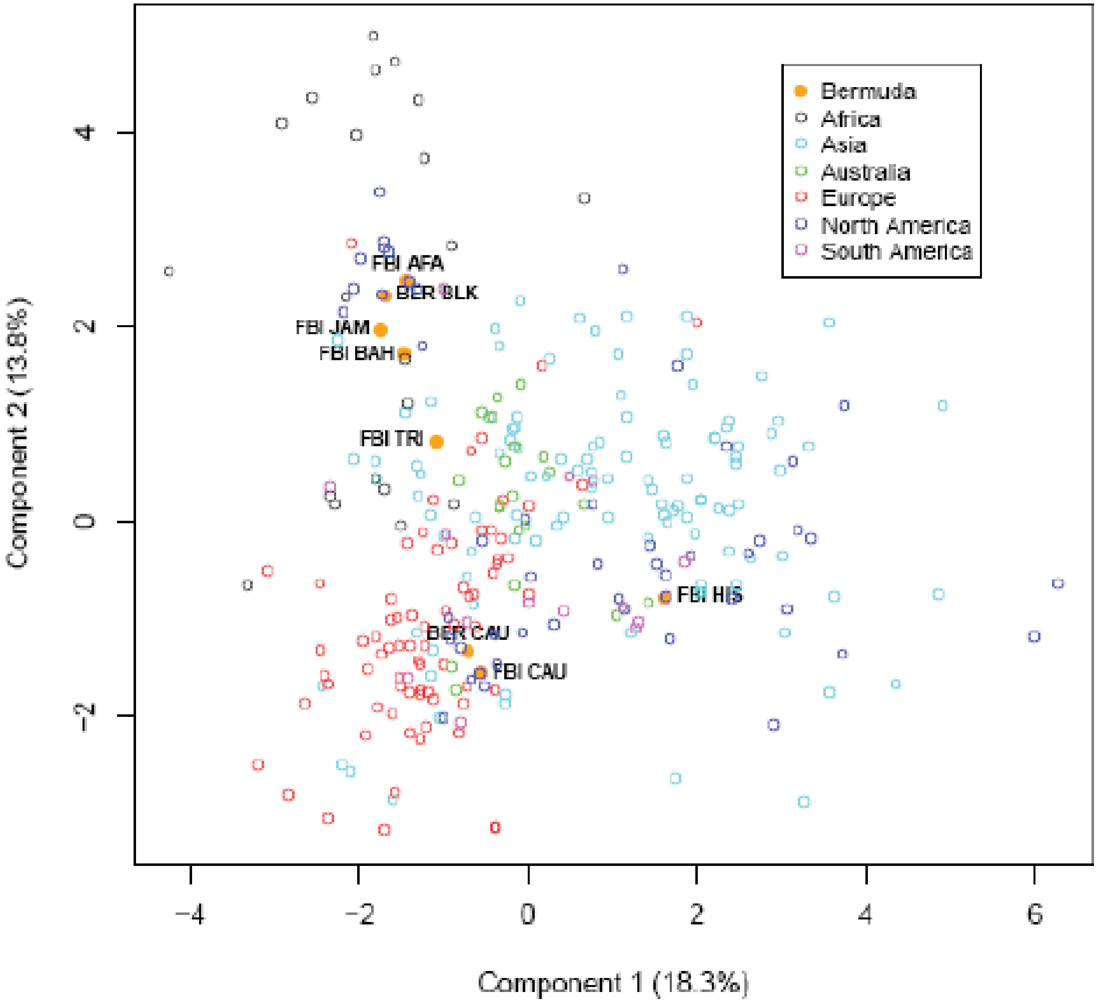
Matching probabilities should be bigger within populations, and more similar among populations that are closer together in time.

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: $\hat{\beta}_{12} = ([\tilde{M}_1 + \tilde{M}_2]/2 - \tilde{M}_{12})/(1 - \tilde{M}_{12})$.

$\tilde{M}_1$: 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
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
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

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:

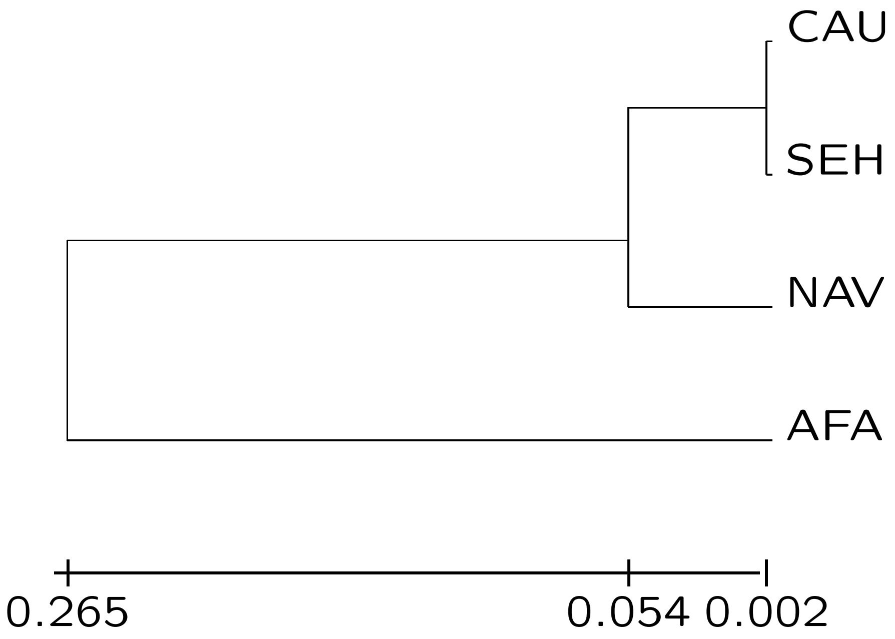
AFA distance $= (0.303+0.254)/2 = 0.278$
NAV distance $= (0.054+0.054)/2 = 0.054$

The new distance matrix is

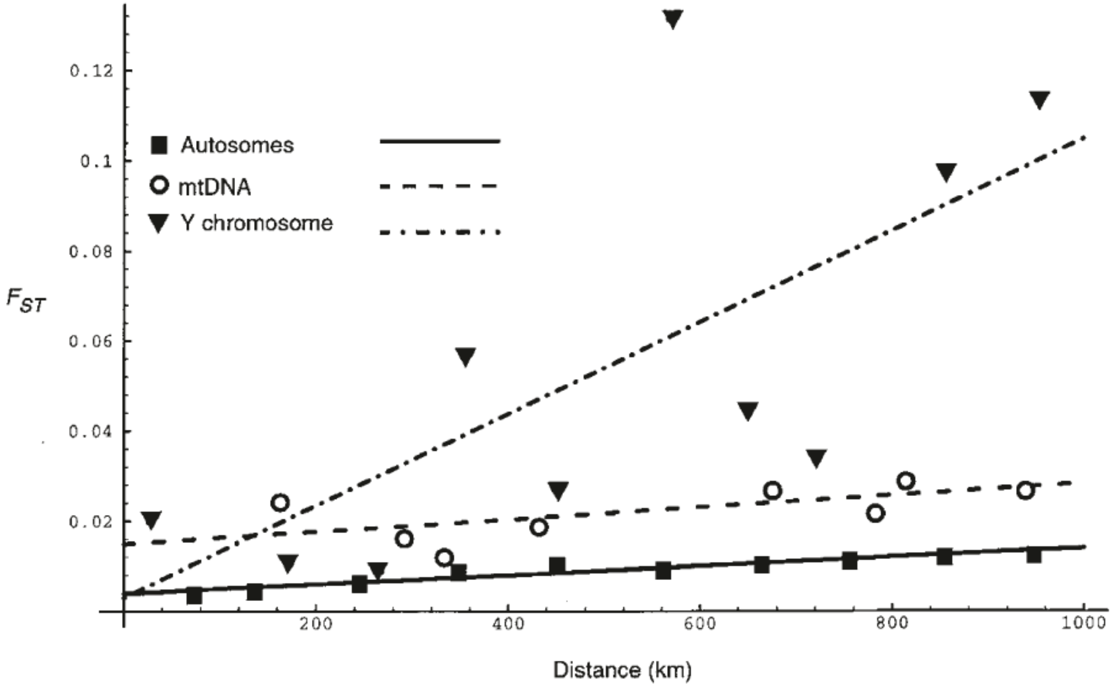
	AFA	CAU/SEH	NAV
AFA	–		
CAU/SEH	0.278	–	
NAV	0.242	0.054	–

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.]

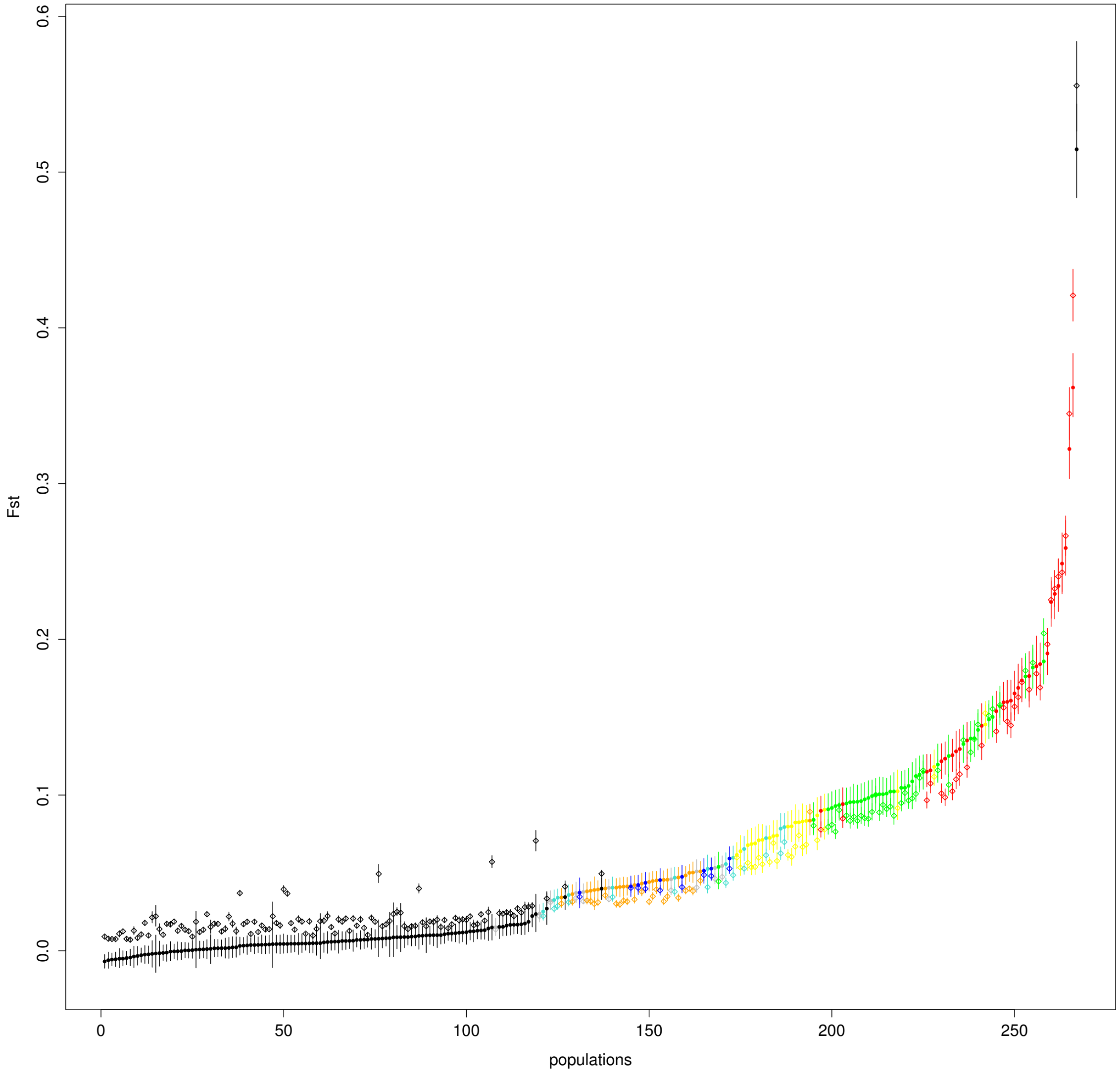
Worldwide Survey of STR Data

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

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.

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

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

For example:

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

Partial Matching

For autosomal markers, two profiles may be:

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

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

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

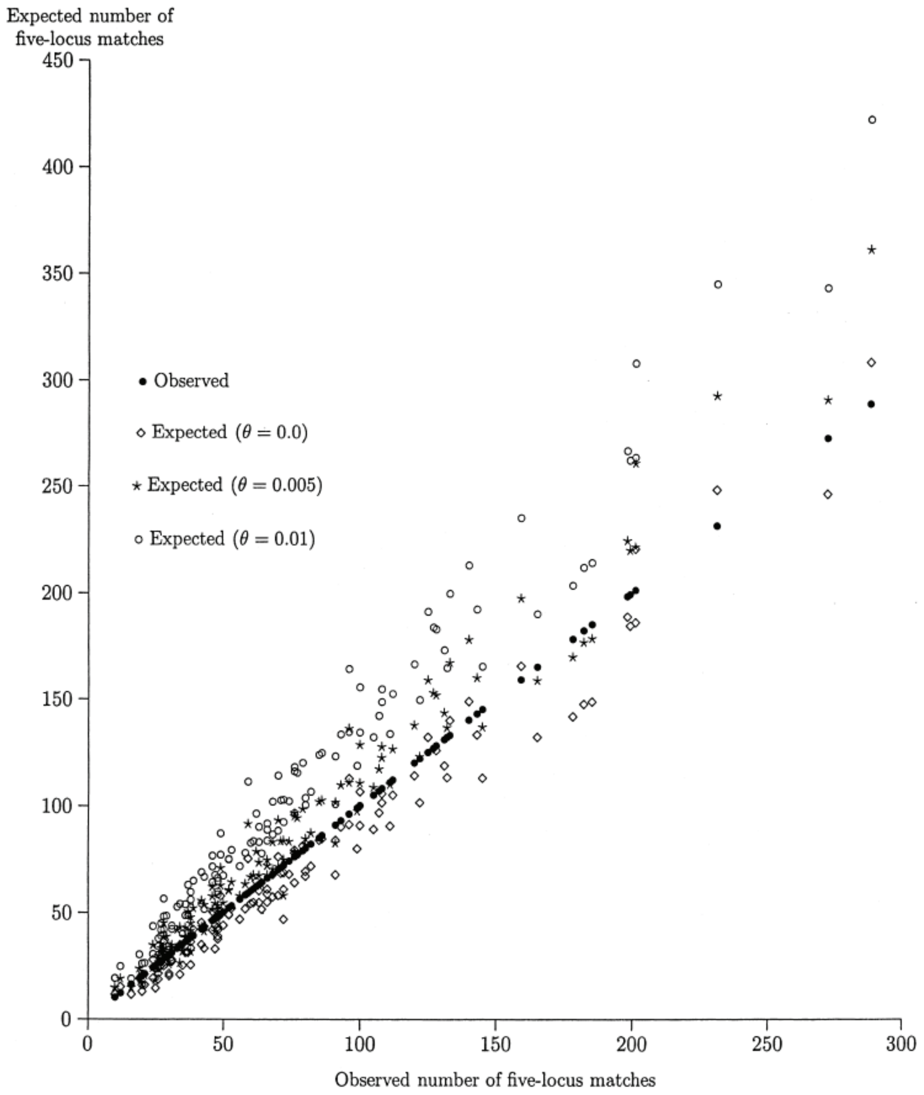
FBI Database Matching Counts

Matching		Number of Partially Matching Loci												
loci	θ	0	1	2	3	4	5	6	7	8	9	10	11	12
0	Obs.	0	3	18	92	249	624	1077	1363	1116	849	379	112	25
	.000	0	2	19	90	293	672	1129	1403	1290	868	415	134	26
	.010	0	2	14	70	236	566	992	1289	1241	875	439	148	30
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: $\widehat{\beta}$ 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.

Predicted Kinship Values

$$\begin{tikzcd}
& & A \arrow[dl] \arrow[dr] & & \\
{} \arrow[dr] & \vdots \arrow[d] & & \vdots \arrow[d] & {} \arrow[dl] \\
& X \arrow[dr] & & Y \arrow[dl] & \\
& & I & &
\end{tikzcd}$$

Identify the path linking the parents X, Y of I to their common ancestor(s).

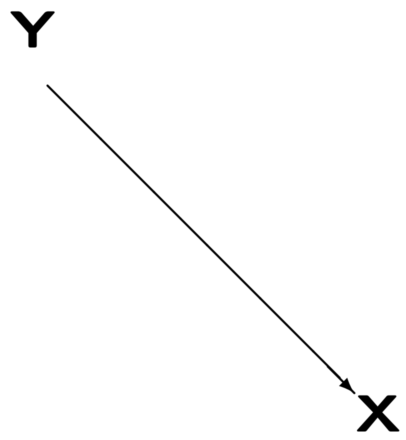
Path Counting

If the parents X, Y of an individual I have ancestor A in common, and if there are n individuals (including X, Y, I) in the path linking the parents through A, then the inbreeding coefficient of I, or the kinship of X and Y, is

$$F_I = \theta_{XY} \quad = \quad \left(\frac{1}{2}\right)^n (1 + F_A)$$

If there are several ancestors, this expression is summed over all the ancestors.

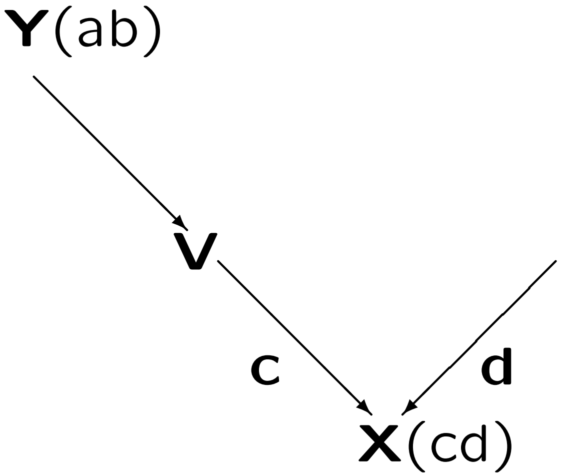
Parent-Child



The common ancestor of parent X and child Y is X. The path linking X, Y to their common ancestor is YX and this has $n = 2$ individuals. Therefore

$$\theta_{XY} = \left(\frac{1}{2}\right)^2 = \frac{1}{4}$$

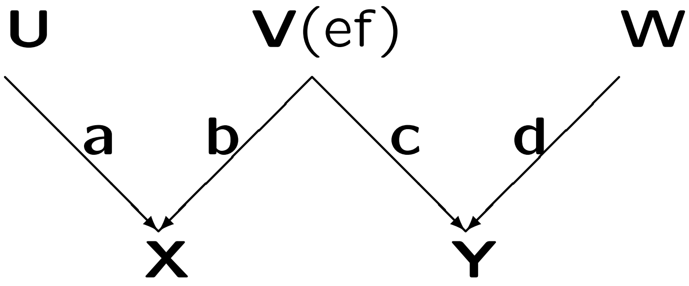
Grandparent-grandchild



The path joining X to Y is XVY with $n = 3$:

$$\theta_{XY} = \left(\frac{1}{2}\right)^3 = \frac{1}{8}$$

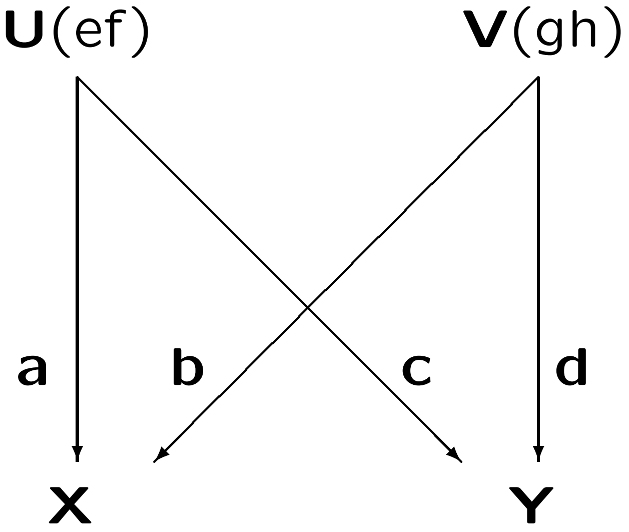
Half sibs



There is one path joining X to Y: XVY with $n = 3$:

$$\theta_{XY} \; = \; \left(\frac{1}{2}\right)^3 = \frac{1}{8}$$

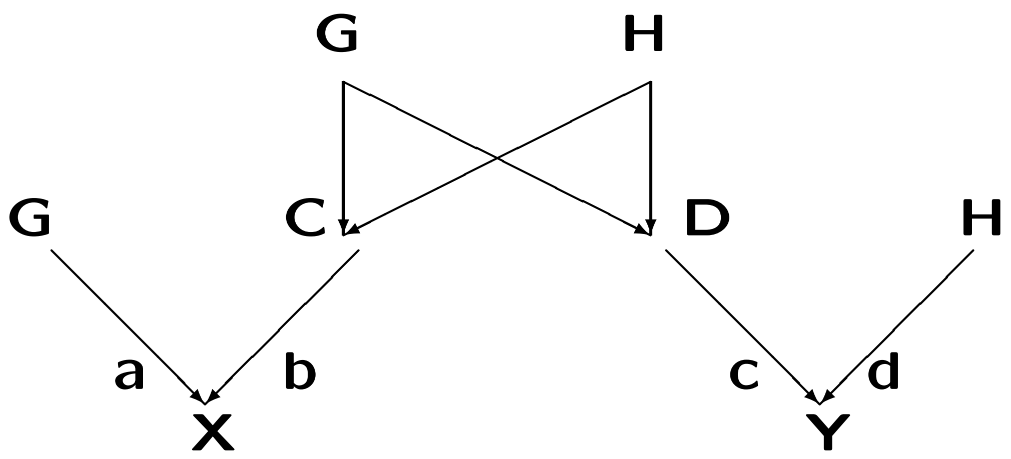
Full sibs



There are two paths joining X to Y: XUY and XVY each with $n = 3$:

$$\theta_{XY} \;=\; \left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^3 = \frac{1}{4}$$

First cousins



Common Relatives

Relationship	Kinship
Identical Twins	0.5
Parent Child	0.25
Full Sibs	0.25
Half Sibs	0.125
Double First Cousins	0.125
First Cousins	0.0625
Uncle Niece	0.0625
Unrelated	0

Comparing Hypothesized Relationships

Current practise is to compare the likelihoods of two profiles under alternative hypotheses about their degrees of relatedness.

On the verge now of being able to estimate the degree of relatedness, especially with very large numbers of SNP markers.

Estimating Kinship

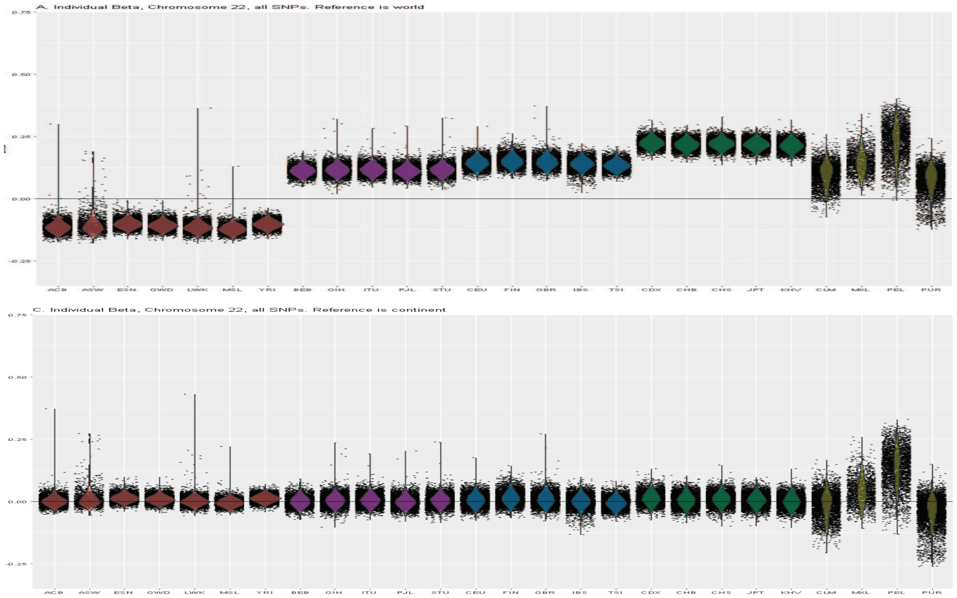
The proportion $\tilde{M}_{XY}$ of pairs of alleles, one from individual X and one from individual Y, that match is 0, 0.5 or 1:
Proportion=1: AA and AA
Proportion=0.5: AA and AB or AB and AB
Proportion=0: AA and BB or AA and BC or AB and CD

Averaging over all pairs of individuals, one per population, the observed proportion is $\tilde{M}^B$. The kinship of individuals X, Y, relative to that of all individuals in different populations is

$$\hat{\theta}_{XY} = \frac{\tilde{M}_{XY} - \tilde{M}^B}{1 - \tilde{M}^B}$$

Kinship is relative, not absolute

Top row: Whole world reference. Bottom row: Continental group reference.



Beta estimates

Chromosome 22 data from 1000 Genomes.

Continents (left to right): AFR, SAS, EUR, EAS, AMR

Populations (l to r):**AFR**: ACB, ASW, ESN, GWD, LWK, MSL, YRI;
SAS: BEB, GIH, ITU, PJL, STU; **EUR**: CEU, FIN, GBR, IBS, TSI;
EAS: CDX, CHB, CHS, JPT; **AMR**: KHV, CLM, MXL, PEL, PUR

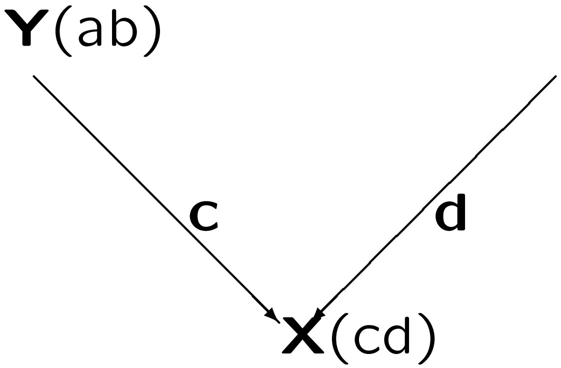
k-coefficients

The coancestry coefficient is the probability of a pair of alleles being ibd.

For joint genotypic frequencies, and for a more detailed characterization of relatedness of two non-inbred individuals, we need the probabilities that they carry 0, 1, or 2 pairs of ibd alleles. For example: their two maternal alleles may be ibd or not ibd, and their two paternal alleles may be ibd or not.

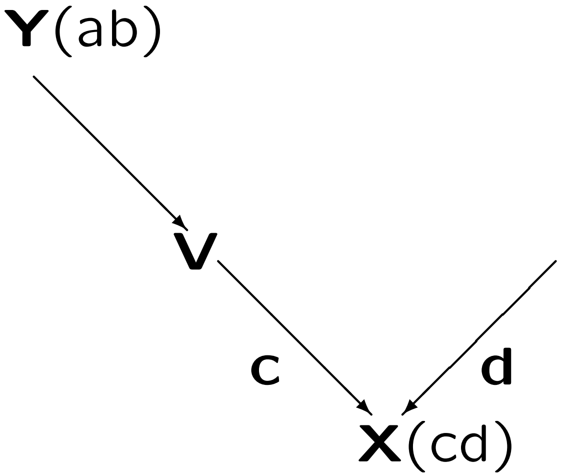
The probabilities of two individuals having 0, 1 or 2 pairs of ibd alleles are written as k_0, k_1, k_2 and $\theta = \frac{1}{2}k_2 + \frac{1}{4}k_1$.

Parent-Child



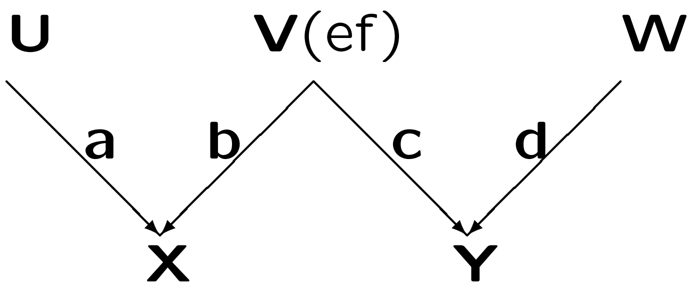
$$\text{Pr}(c \equiv a) = 0.5, \quad \text{Pr}(c \equiv b) = 0.5, \quad k_1 = 1$$

Grandparent-grandchild



$$\Pr(c \equiv a) = 0.25, \quad \Pr(c \equiv b) = 0.25, \quad k_1 = 0.5 \& k_0 = 0.5$$

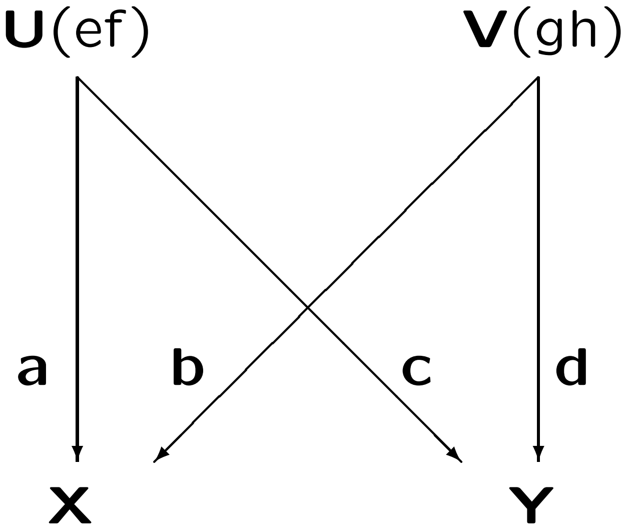
Half sibs



		0.5	0.5
		$c \equiv e$	$c \equiv f$
0.5	$b \equiv e$	0.25	0.25
0.5	$b \equiv f$	0.25	0.25

Therefore $k_1 = 0.5$ so $k_0 = 0.5$.

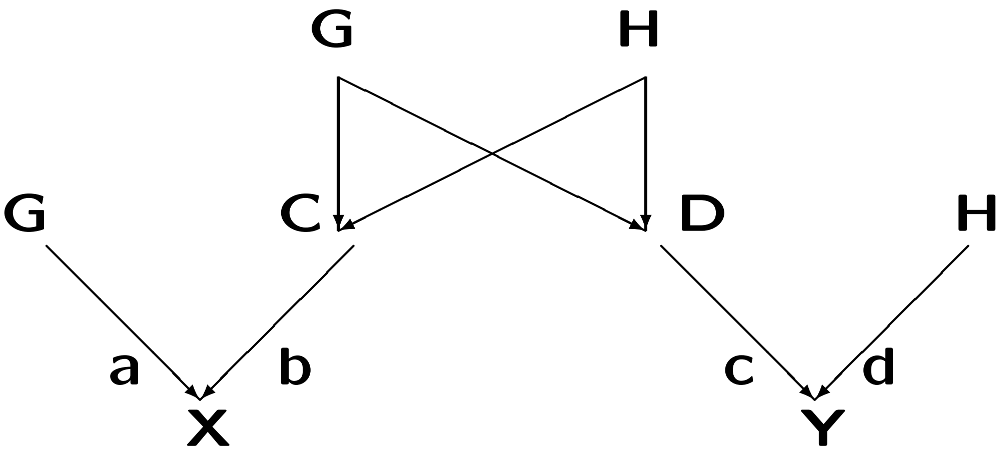
Full sibs



		0.5	0.5
		$b \equiv d$	$b \not\equiv d$
0.5	$a \equiv c$	0.25	0.25
0.5	$a \not\equiv c$	0.25	0.25

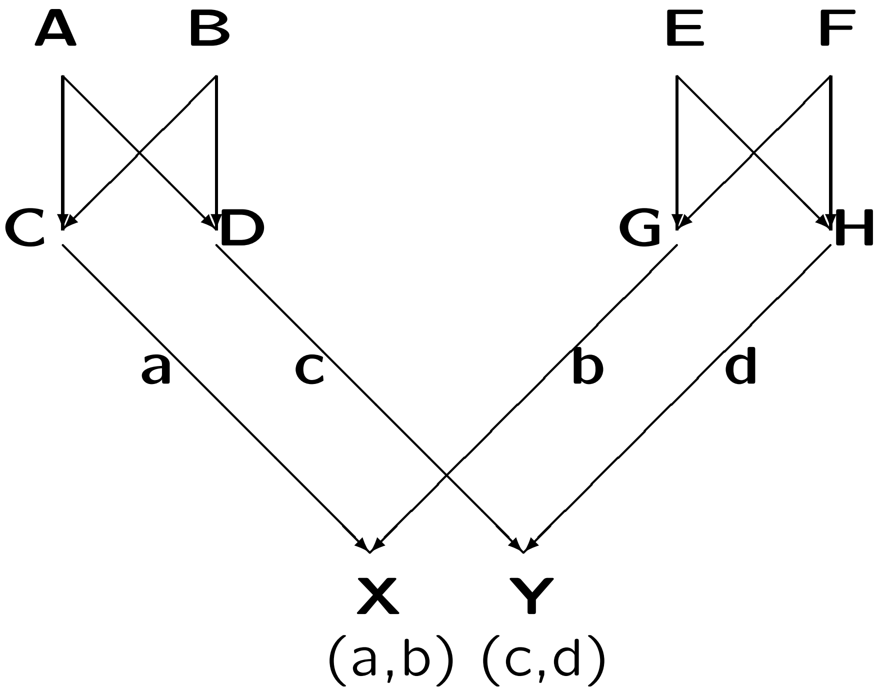
$$k_0 = 0.25, k_1 = 0.50, k_2 = 0.25$$

First cousins



Double First Cousins

What are the k's for double first cousins?

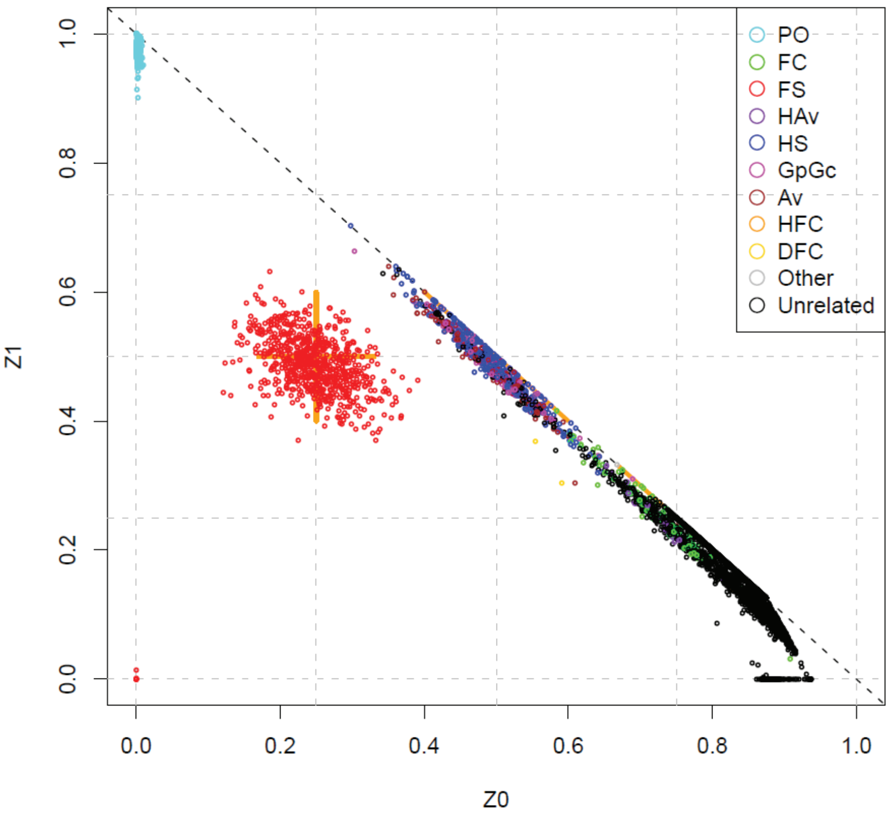


Non-inbred Relatives

Relationship	k_2	k_1	k_0	$\theta = \frac{1}{2}k_2 + \frac{1}{4}k_1$
Identical twins	1	0	0	$\frac{1}{2}$
Full sibs	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$
Parent-child	0	1	0	$\frac{1}{4}$
Double first cousins	$\frac{1}{16}$	$\frac{3}{8}$	$\frac{9}{16}$	$\frac{1}{8}$
Half sibs*	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{8}$
First cousins	0	$\frac{1}{4}$	$\frac{3}{4}$	$\frac{1}{16}$
Unrelated	0	0	1	0

* Also grandparent-grandchild and avuncular (e.g. uncle-niece).

PLINK Example



Joint genotypic probabilities

Genotypes	Probability
ii, ii	$k_2p_i^2 + k_1p_i^3 + k_0p_i^4$
ii, jj	$k_0p_i^2p_j^2$
ii, ij	$k_1p_i^2p_j + 2k_0p_i^3p_j$
ii, jk	$2k_0p_i^2p_jp_k$
ij, ij	$2k_2p_ip_j + k_1p_ip_j(p_i + p_j)$ $+ 4k_0p_i^2p_j^2$
ij, ik	$k_1p_ip_jp_k + 4k_0p_i^2p_jp_k$
ij, kl	$4k_0p_ip_jp_kp_l$

Example: Non-inbred full sibs

Genotypes	Probability
ii, ii	$p_i^2(1+p_i)^2/4$
ii, jj	$p_i^2 p_j^2/4$
ii, ij	$p_i p_j(p_i+p_j)/2$
ii, jk	$p_i^2 p_j p_k/2$
ij, ij	$p_i p_j(1+p_i+p_j+2p_i p_j)/2$
ij, ik	$p_i p_j p_k(1+2p_i)/2$
ij, kl	$p_i p_j p_k p_l$

Match Probabilities with θ for Relatives

$$\begin{aligned}
\Pr(\text{Match}) &= k_2 + k_1[\sum_i \Pr(A_iA_iA_i) + \sum_i \sum_{j\neq i} \Pr(A_iA_jA_j)] \\
&\quad + k_0P_2 \\
&= k_2 + k_1[\theta + (1-\theta)S_2] + k_0P_2
\end{aligned}$$

$$\begin{aligned}
\Pr(\text{Partial Match}) &= k_1[2\sum_i \sum_{j\neq i} \Pr(A_iA_iA_j) + \sum_i \sum_{j\neq i} \sum_{k\neq i,j} \Pr(A_iA_jA_k)] \\
&\quad + k_0P_1 \\
&= k_1(1-\theta)(1-S_2) + k_0P_1
\end{aligned}$$

$$\Pr(\text{Mismatch}) = k_0P_0$$

Quantities P_0, P_1, P_2 are given on Slide 29.

Match probabilities with $\theta = 0.03$

Locus	Not related	First-cousins	Parent -child	Full-sibs
D3S1358	.089	.124	.229	.387
vWA	.077	.111	.213	.376
FGA	.048	.078	.166	.345
D8S1179	.083	.119	.227	.384
D21S11	.051	.081	.172	.349
D18S51	.040	.068	.150	.335
D5S818	.175	.216	.339	.463
D13S317	.101	.139	.252	.401
D7S820	.080	.115	.219	.379
CSF1PO	.134	.173	.288	.428
TPOX	.216	.261	.397	.503
THO1	.096	.133	.241	.395
D16S539	.105	.143	.256	.404
Total	2×10^{-14}	2×10^{-12}	6×10^{-9}	5×10^{-6}

Arizona Matches: Mueller Analysis

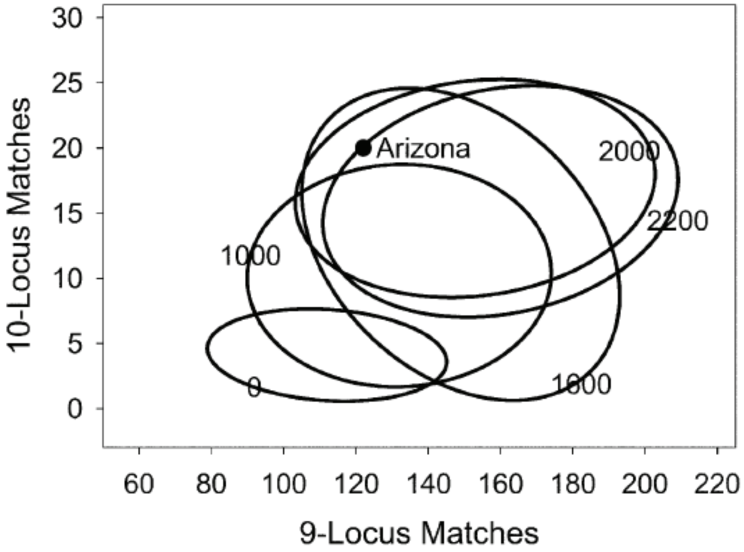


Figure 8. 95% confidence ellipsoids for simulations in which θ was set to 0.015 and the number of full sibs varied. The number on each ellipsoid corresponds to the number of pairs of sibs present in the simulated databases.

Mueller LD. 2008. Journal of Genetics 87:101-107.

Mueller Comment

"The product rule with some minor modification is the most common method for computing the frequency of DNA profiles in forensic laboratories. This method relies critically on the assumption that there is statistical independence between loci. The empirical support for this method comes mainly from tests of independence between pairs of loci (Budowle et al. 1999). However, recent research on finite populations, with mutation and a monogamous mating system shows that departures from the product rule get worse as one looks at more loci (Dr Yun Song, personal communication). Thus, rigorous testing of the product rule predictions at many loci may yield different results than prior work at only two loci. Perhaps the most important qu1ality control issue in forensic DNA typing is determining the adequacy of the methods for computing profile frequencies."

Mueller LD. 2008. Journal of Genetics 87:101-107.

“RELPAIR” calculations

This approach compares the probabilities of two genotypes under alternative hypotheses; H_0: the individuals have a specified relationship, versus H_1: the individuals are unrelated. The alternative is that $k_0 = 1, k_1 = k_2 = 0$ so the likelihood ratios for the two hypotheses are:

$$\begin{aligned}
\mathrm{LR}(MM, MM) &= k_0 + k_1/p_M + k_2/p_M^2 \\
\mathrm{LR}(mm, mm) &= k_0 + k_1/p_m + k_2/p_m^2 \\
\mathrm{LR}(Mm, Mm) &= k_0 + k_1/(4p_M p_m) + k_2/(2p_M p_m)
\end{aligned}$$

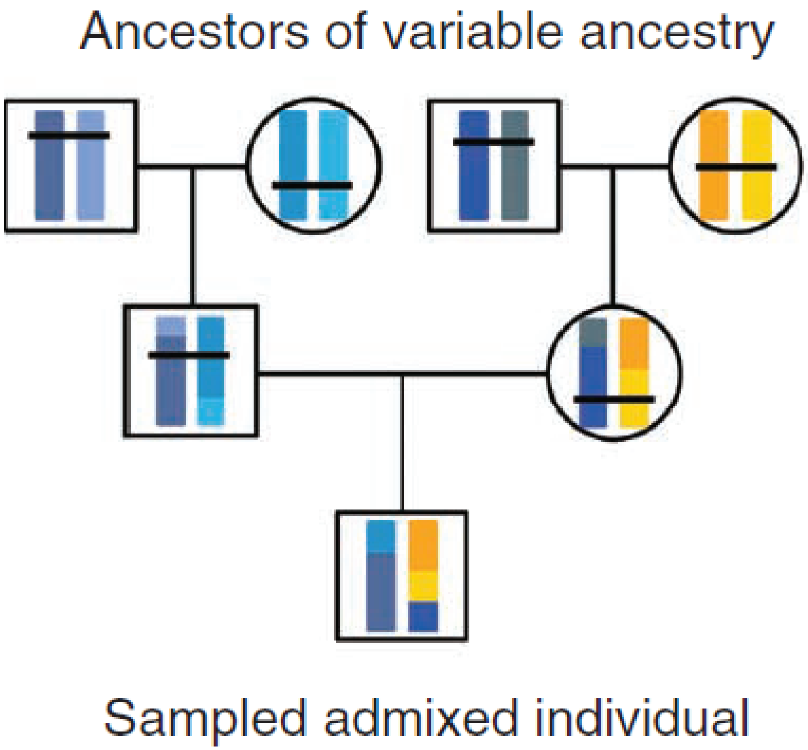
$$\begin{aligned}
\mathrm{LR}(MM, Mm) &= k_0 + k_1/(2p_M) \\
\mathrm{LR}(mm, Mm) &= k_0 + k_1/(2p_m)
\end{aligned}$$

$$\mathrm{LR}(MM, mm) = k_0$$

Forensic Genealogy

Recombination

One Morgan is the length along a chromosome in which 1 recombination event is expected to occur. The human genome has a total map length of 36M, meaning that each chromosome is expected to have 1-2 recombination events per generation. A centi-Morgan (cM) is one-hundreth of a Morgan.



Wegmann D et al. 2011. Nature Genetics 43:84

The Shared cM Project

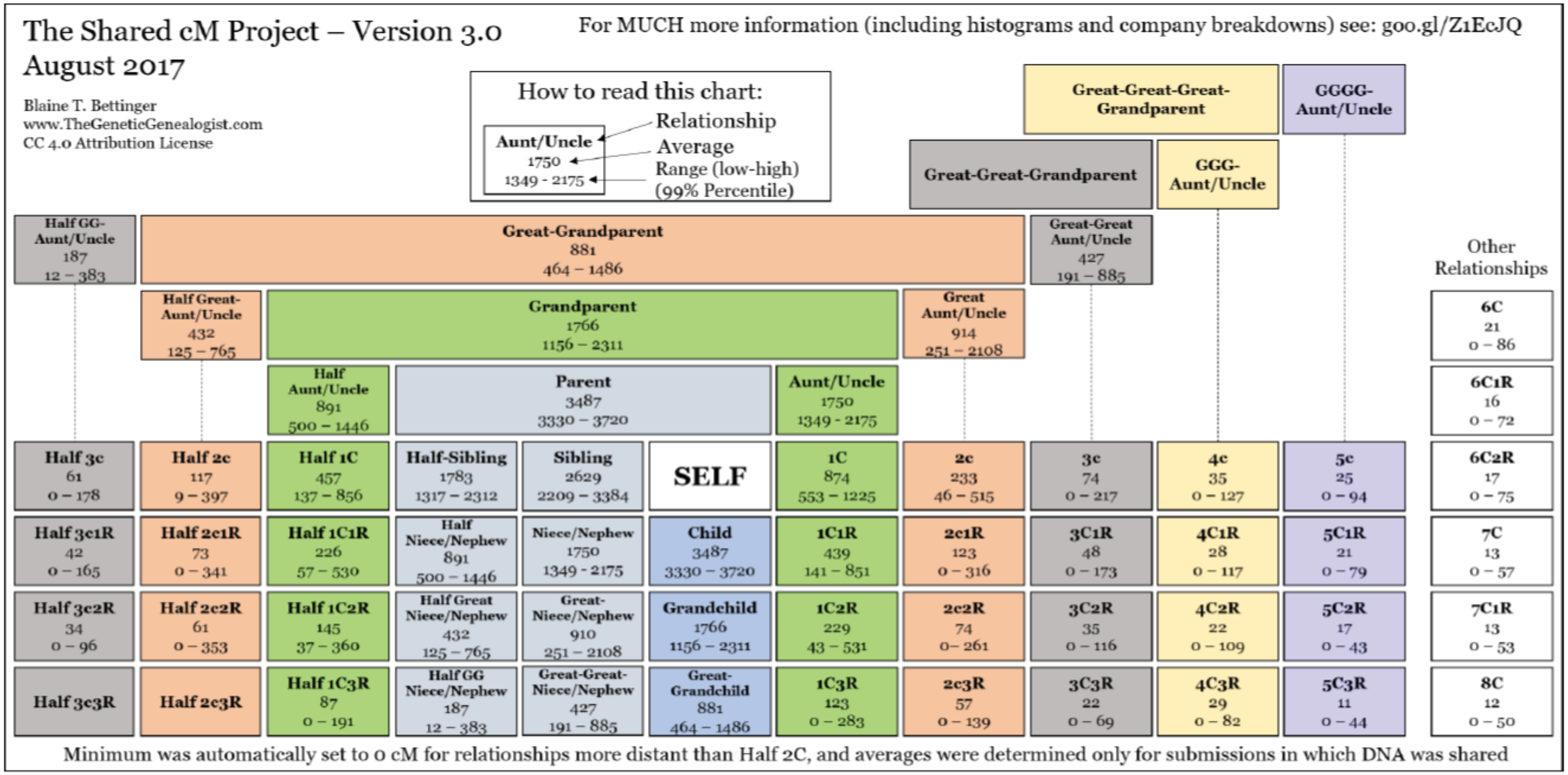
https://thegeneticgenealogist.com/

https://thegeneticgenealogist.com/2017/08/26/august-2017-update-to-the-shared-cm-project/

The Shared cM Project

The Shared cM Project – Version 3.0 (August 2017)

Figure 1. The Relationship Chart



The Shared cM Project – Version 3.0
August 2017

Blaine T. Bettinger
www.TheGeneticGenealogist.com
CC 4.0 Attribution License

For MUCH more information (including histograms and company breakdowns) see: goo.gl/Z1EcJQ

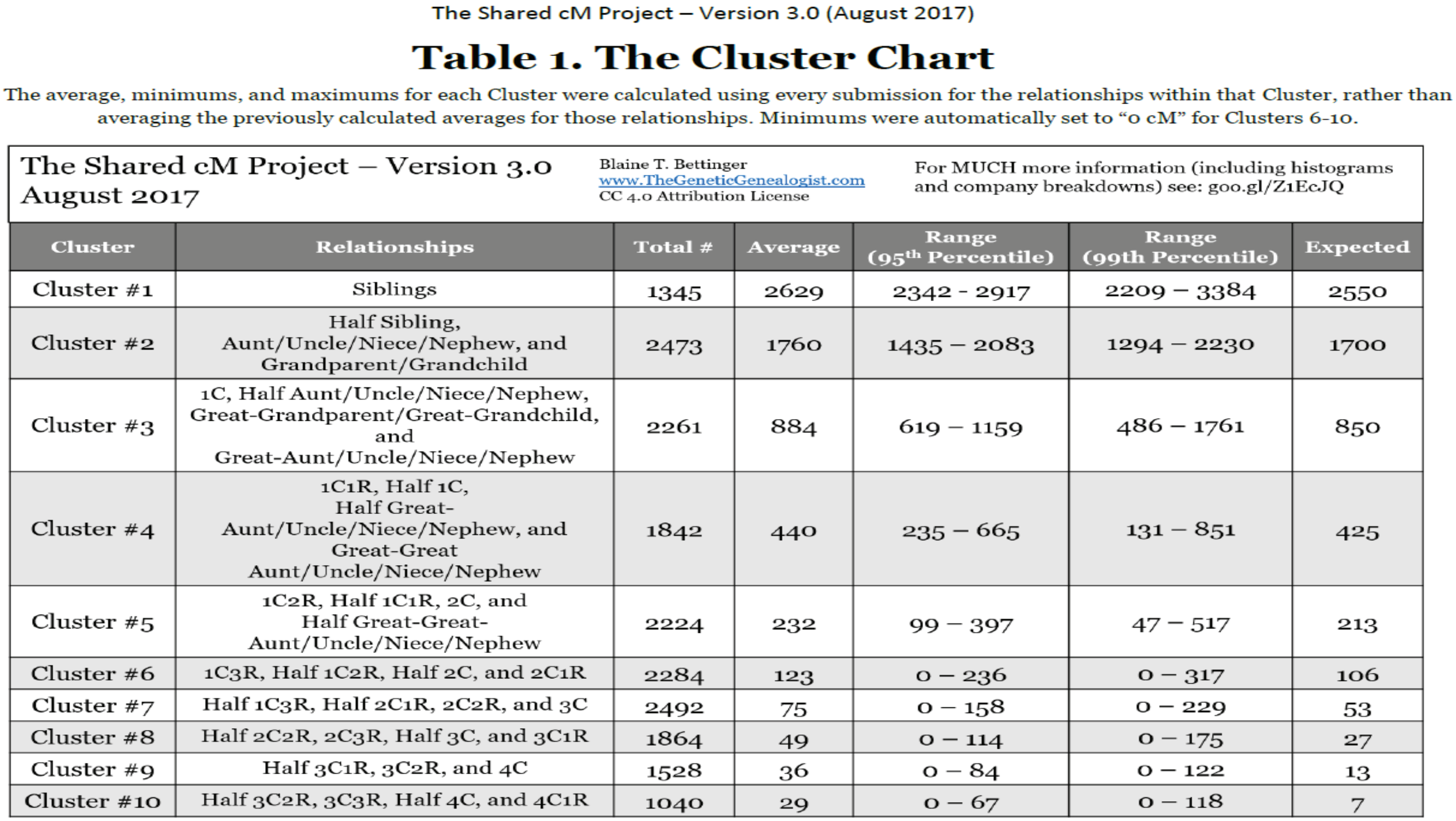
How to read this chart:

Aunt/Uncle – Relationship
1750 – Average
1349 - 2175 – Range (low-high) (99% Percentile)

								Great-Great-Great-Grandparent		GGGG-Aunt/Uncle	
							Great-Great-Grandparent		GGG-Aunt/Uncle		
Half GG-Aunt/Uncle 187 12 – 383	Great-Grandparent 881 464 – 1486							Great-Great Aunt/Uncle 427 191 – 885			Other Relationships
	Half Great-Aunt/Uncle 432 125 – 765	Grandparent 1766 1156 – 2311					Great Aunt/Uncle 914 251 – 2108				6C 21 0 – 86
		Half Aunt/Uncle 891 500 – 1446	Parent 3487 3330 – 3720			Aunt/Uncle 1750 1349 - 2175					6C1R 16 0 – 72
Half 3c 61 0 – 178	Half 2c 117 9 – 397	Half 1C 457 137 – 856	Half-Sibling 1783 1317 – 2312	Sibling 2629 2209 – 3384	SELF	1C 874 553 – 1225	2c 233 46 – 515	3c 74 0 – 217	4c 35 0 – 127	5c 25 0 – 94	6C2R 17 0 – 75
Half 3c1R 42 0 – 165	Half 2c1R 73 0 – 341	Half 1C1R 226 57 – 530	Half Niece/Nephew 891 500 – 1446	Niece/Nephew 1750 1349 - 2175	Child 3487 3330 – 3720	1C1R 439 141 – 851	2c1R 123 0 – 316	3C1R 48 0 – 173	4C1R 28 0 – 117	5C1R 21 0 – 79	7C 13 0 – 57
Half 3c2R 34 0 – 96	Half 2c2R 61 0 – 353	Half 1C2R 145 37 – 360	Half Great Niece/Nephew 432 125 – 765	Great-Niece/Nephew 910 251 – 2108	Grandchild 1766 1156 – 2311	1C2R 229 43 – 531	2c2R 74 0– 261	3C2R 35 0 – 116	4C2R 22 0 – 109	5C2R 17 0 – 43	7C1R 13 0 – 53
Half 3c3R	Half 2c3R	Half 1C3R 87 0 – 191	Half GG Niece/Nephew 187 12 – 383	Great-Great-Niece/Nephew 427 191 – 885	Great-Grandchild 881 464 – 1486	1C3R 123 0 – 283	2c3R 57 0 – 139	3C3R 22 0 – 69	4C3R 29 0 – 82	5C3R 11 0 – 44	8C 12 0 – 50

Minimum was automatically set to 0 cM for relationships more distant than Half 2C, and averages were determined only for submissions in which DNA was shared

The Shared cM Project



The Shared cM Project – Version 3.0 (August 2017)

Table 1. The Cluster Chart

The average, minimums, and maximums for each Cluster were calculated using every submission for the relationships within that Cluster, rather than averaging the previously calculated averages for those relationships. Minimums were automatically set to “0 cM” for Clusters 6-10.

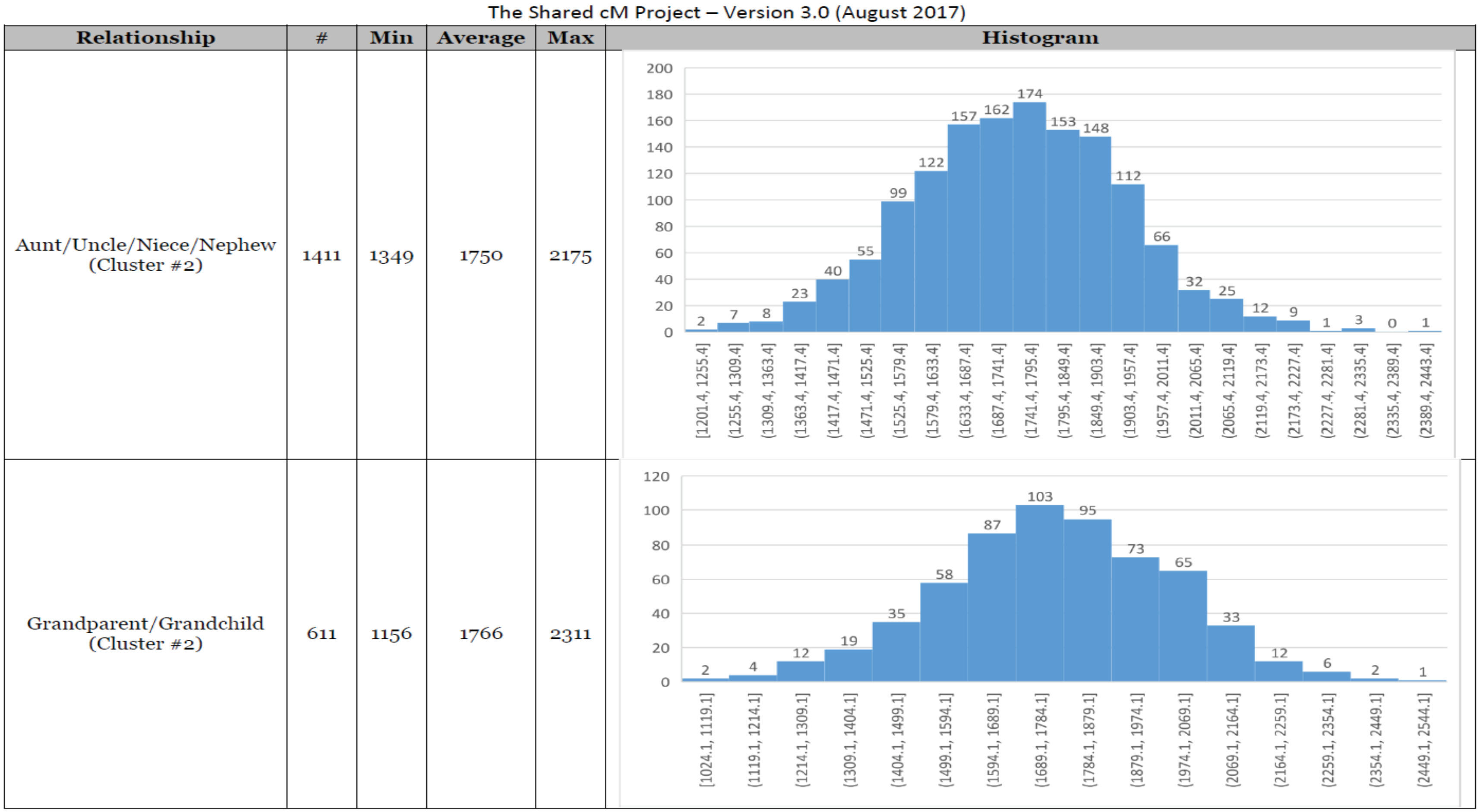
The Shared cM Project – Version 3.0
August 2017

Blaine T. Bettinger
www.TheGeneticGenealogist.com
CC 4.0 Attribution License

For MUCH more information (including histograms and company breakdowns) see: goo.gl/Z1EcJQ

Cluster	Relationships	Total #	Average	Range (95th Percentile)	Range (99th Percentile)	Expected
Cluster #1	Siblings	1345	2629	2342 - 2917	2209 – 3384	2550
Cluster #2	Half Sibling, Aunt/Uncle/Niece/Nephew, and Grandparent/Grandchild	2473	1760	1435 – 2083	1294 – 2230	1700
Cluster #3	1C, Half Aunt/Uncle/Niece/Nephew, Great-Grandparent/Great-Grandchild, and Great-Aunt/Uncle/Niece/Nephew	2261	884	619 – 1159	486 – 1761	850
Cluster #4	1C1R, Half 1C, Half Great-Aunt/Uncle/Niece/Nephew, and Great-Great Aunt/Uncle/Niece/Nephew	1842	440	235 – 665	131 – 851	425
Cluster #5	1C2R, Half 1C1R, 2C, and Half Great-Great-Aunt/Uncle/Niece/Nephew	2224	232	99 – 397	47 – 517	213
Cluster #6	1C3R, Half 1C2R, Half 2C, and 2C1R	2284	123	0 – 236	0 – 317	106
Cluster #7	Half 1C3R, Half 2C1R, 2C2R, and 3C	2492	75	0 – 158	0 – 229	53
Cluster #8	Half 2C2R, 2C3R, Half 3C, and 3C1R	1864	49	0 – 114	0 – 175	27
Cluster #9	Half 3C1R, 3C2R, and 4C	1528	36	0 – 84	0 – 122	13
Cluster #10	Half 3C2R, 3C3R, Half 4C, and 4C1R	1040	29	0 – 67	0 – 118	7

The Shared cM Project



The Shared cM Project – Version 3.0 (August 2017)

Relationship	#	Min	Average	Max	Histogram
Aunt/Uncle/Niece/Nephew (Cluster #2)	1411	1349	1750	2175	
Grandparent/Grandchild (Cluster #2)	611	1156	1766	2311	

200
180
160
140
120
100
80
60
40
20
0
2
7
8
23
40
55
99
122
157
162
174
153
148
112
66
32
25
12
9
1
3
0
1
[1201.4, 1255.4]
(1255.4, 1309.4]
(1309.4, 1363.4]
(1363.4, 1417.4]
(1417.4, 1471.4]
(1471.4, 1525.4]
(1525.4, 1579.4]
(1579.4, 1633.4]
(1633.4, 1687.4]
(1687.4, 1741.4]
(1741.4, 1795.4]
(1795.4, 1849.4]
(1849.4, 1903.4]
(1903.4, 1957.4]
(1957.4, 2011.4]
(2011.4, 2065.4]
(2065.4, 2119.4]
(2119.4, 2173.4]
(2173.4, 2227.4]
(2227.4, 2281.4]
(2281.4, 2335.4]
(2335.4, 2389.4]
(2389.4, 2443.4]
120
100
80
60
40
20
0
2
4
12
19
35
58
87
103
95
73
65
33
12
6
2
1
[1024.1, 1119.1]
(1119.1, 1214.1]
(1214.1, 1309.1]
(1309.1, 1404.1]
(1404.1, 1499.1]
(1499.1, 1594.1]
(1594.1, 1689.1]
(1689.1, 1784.1]
(1784.1, 1879.1]
(1879.1, 1974.1]
(1974.1, 2069.1]
(2069.1, 2164.1]
(2164.1, 2259.1]
(2259.1, 2354.1]
(2354.1, 2449.1]
(2449.1, 2544.1]

Henn et al., 2012

“To infer identity by descent, we scanned each pair of genomes for long runs of genotype pairs that lack opposite homozygotes. We define inferred IBDhalf as the sum of the lengths of genomic segments where two individuals share DNA identical by state for at least one of the homologous chromosomes. This method is computationally feasible in large sample sets .”

Henn BL, Hon L, Macpherson JM, Eriksson N, Saxonov S, Pe’er I, Mountain JL. 2012. Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. PLoS One 7:e34267.

Henn et al., 2012

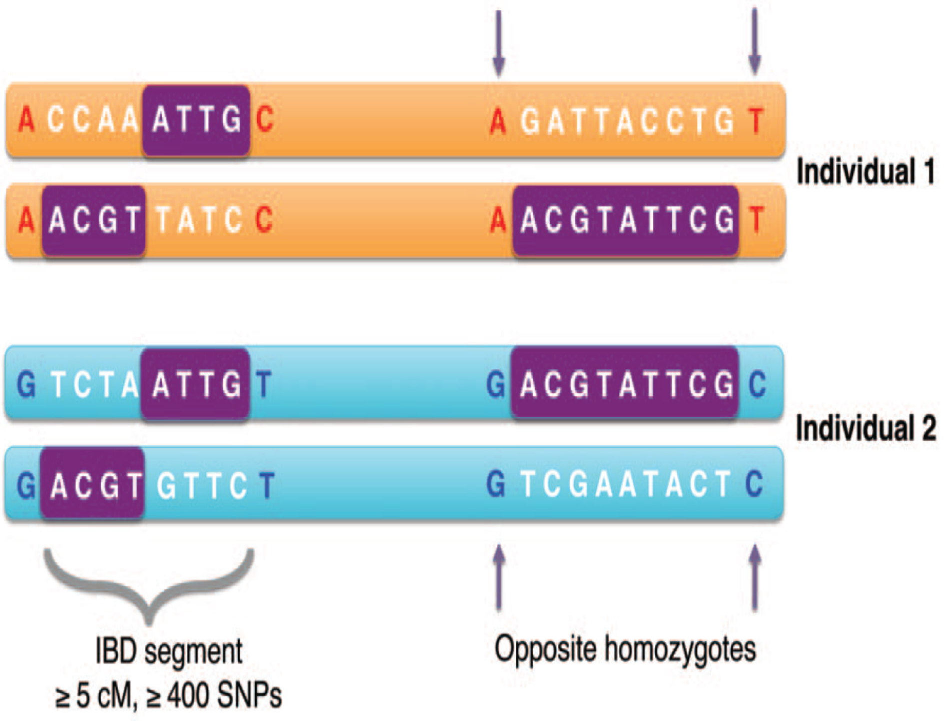


Figure 1. Schematic of IBD$_{half}$ inference method. IBD$_{half}$ segments were inferred from unphased genotype data where a series of alleles were identical by state for *at least one* of the homologous chromosomes in a given pair of individuals. IBD segments are indicated in purple. The boundaries of the IBD segments are defined by "opposite homozygotes". Additionally, an IBD region had to be minimally 5 cM in length and contains >400 genotyped SNPs that were homozygous in at least one of the two individuals being compared (see *Methods*).

Henn et al., 2012

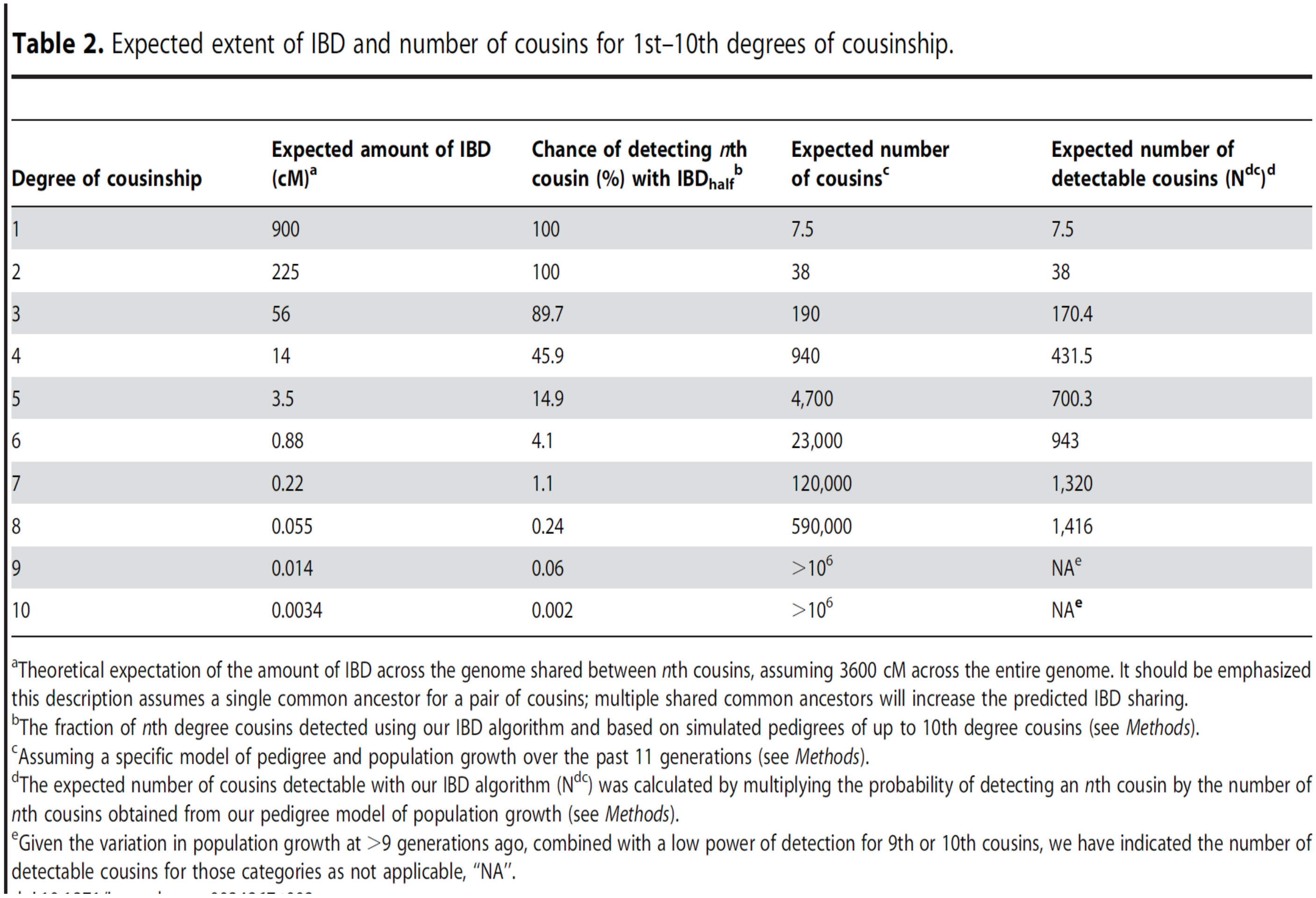


Table 2. Expected extent of IBD and number of cousins for 1st–10th degrees of cousinship.

Degree of cousinship	Expected amount of IBD (cM)[a]	Chance of detecting *n*th cousin (%) with IBD_{half}[b]	Expected number of cousins[c]	Expected number of detectable cousins (N^{dc})[d]
1	900	100	7.5	7.5
2	225	100	38	38
3	56	89.7	190	170.4
4	14	45.9	940	431.5
5	3.5	14.9	4,700	700.3
6	0.88	4.1	23,000	943
7	0.22	1.1	120,000	1,320
8	0.055	0.24	590,000	1,416
9	0.014	0.06	$>10^6$	NA[e]
10	0.0034	0.002	$>10^6$	NA[e]

[a] Theoretical expectation of the amount of IBD across the genome shared between *n*th cousins, assuming 3600 cM across the entire genome. It should be emphasized this description assumes a single common ancestor for a pair of cousins; multiple shared common ancestors will increase the predicted IBD sharing.

[b] The fraction of *n*th degree cousins detected using our IBD algorithm and based on simulated pedigrees of up to 10th degree cousins (see *Methods*).

[c] Assuming a specific model of pedigree and population growth over the past 11 generations (see *Methods*).

[d] The expected number of cousins detectable with our IBD algorithm (N^{dc}) was calculated by multiplying the probability of detecting an *n*th cousin by the number of *n*th cousins obtained from our pedigree model of population growth (see *Methods*).

[e] Given the variation in population growth at >9 generations ago, combined with a low power of detection for 9th or 10th cousins, we have indicated the number of detectable cousins for those categories as not applicable, "NA".

Henn et al., 2012

We inferred that two individuals share DNA IBD from unphased data. We inferred boundaries of IBD by comparing two individuals' genotypes at a locus and identifying SNPs where one individuals genotype is homozygous for one allele and the other individual's genotype is homozygous for a second allele. By characterizing stretches that lacked these opposite homozygotes, we defined regions that contain at least half IBD between two individuals. That is, an IBDhalf segment was characterized by a series of alleles that were identical by state for at least one of the homologous chromosomes in a given pair of individuals. We define IBDhalf as the sum of the lengths of genomic segments where two individuals are inferred to share DNA identical by descent for at least one of the homologous chromosomes.

Henn et al., 2012

We additionally enforced two criteria to increase our confidence that a region represents DNA that is IBD: first, the region is minimally 5 cM in length and second, it contains at least 400 genotyped SNPs that are homozygous in at least one of the two individuals being compared, ensuring that there is both sufficient genotype coverage and genetic distance defining the IBD region. Finally, we accepted a comparison as IBD if the longest segment in the comparison was at least 7 cM.”

Henn et al., 2012

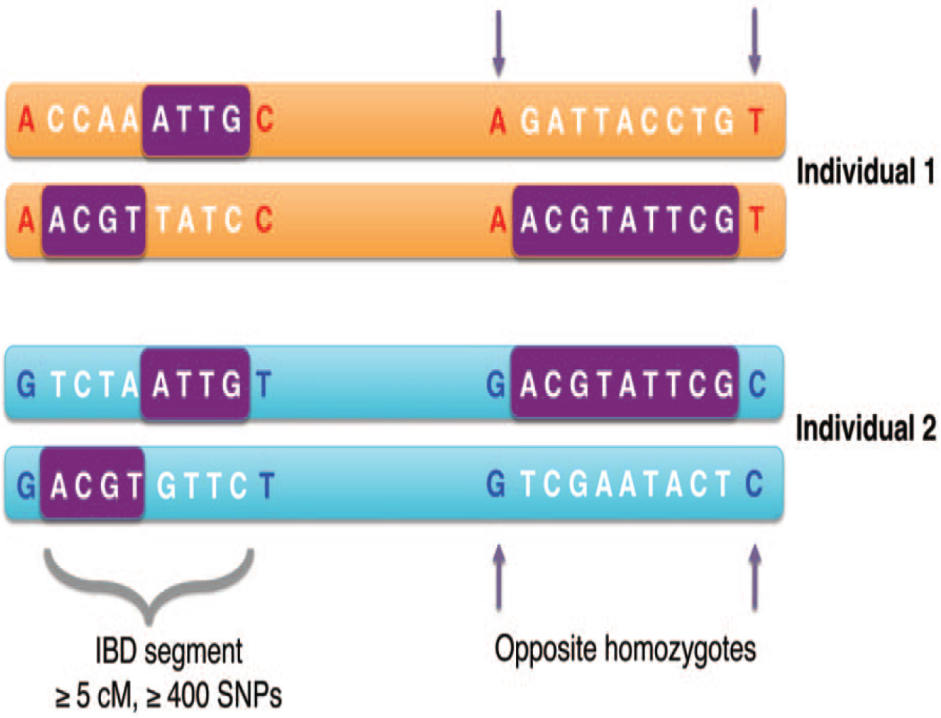

A CCAA ATTG C A GATTACCTG T
A ACGT TATC C A ACGTATTCG T
Individual 1
G TCTA ATTG T G ACGTATTCG C
G ACGT GTTCT G TCGAATACT C
Individual 2
IBD segment
≥ 5 cM, ≥ 400 SNPs
Opposite homozygotes

Figure 1. Schematic of IBD$_{half}$ inference method. IBD$_{half}$ segments were inferred from unphased genotype data where a series of alleles were identical by state for *at least one* of the homologous chromosomes in a given pair of individuals. IBD segments are indicated in purple. The boundaries of the IBD segments are defined by "opposite homozygotes". Additionally, an IBD region had to be minimally 5 cM in length and contains >400 genotyped SNPs that were homozygous in at least one of the two individuals being compared (see *Methods*).

Genealogy Search

Suppose a GEDMatch search for an evidence profile E reveals two first cousins for the source of E: $C1, C2$.

E and $C1$ have two of their four grandparents in common. Think of the four grandparents of $C1$ and trace their descendants $D1$: there are the parents, uncles, aunts and cousins of $C1$.

E and $C2$ have two of their four grandparents in common. Think of the four grandparents of $C2$ and trace their descendants $D2$: there are the parents, uncles, aunts and cousins of $C2$.

The source of E belongs to both $D1$ and $D2$.