OTHER TOPICS

SINGLE CONTRIBUTORS

Profile and Match Probabilities

Profile Probability:

$$\begin{aligned} \Pr(AA) &= p_A^2 + \theta p_A(1-p_A) \leq 2p_A \\ \Pr(AB) &= 2p_Ap_B - 2\theta p_Ap_B \leq 2p_Ap_B \end{aligned}$$

Match Probability:

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

Relatives

$$\begin{aligned}\Pr(AA|AA) &= k_2 + k_1 p_A + k_0 p_A^2 \\ \Pr(AB|AB) &= k_2 + \frac{1}{2}(p_A + p_B)k_1 + 2k_0 p_A p_B\end{aligned}$$

For unilineal relatives, $k_2 = 0, k_1 + k_0 = 1$ and kinship $\theta = k_1/4$:

$$\begin{aligned}\Pr(AA|AA) &= p_A^2 + 4\theta p_A(1 - p_A) \\ \Pr(AB|AB) &= 2p_A p_B + 2\theta(p_A + p_B - 4p_A p_B)\end{aligned}$$

For full-sibs, $k_1 = k_0 = 1/4, k_1 = 1/2$:

$$\begin{aligned}\Pr(AA|AA) &= \frac{1}{4}(1 + p_A)^2 \\ \Pr(AB|AB) &= \frac{1}{4}(1 + p_A + p_B + 2p_A p_B)\end{aligned}$$

Relatives with Population Structure

$$\begin{aligned}\Pr(AA|AA) &= k_2 + k_1\frac{2\theta + (1-\theta)p_A}{1+\theta} \\ &\quad + k_0\frac{[2\theta + (1-\theta)p_A][3\theta + (1-\theta)p_A]}{(1+\theta)(1+2\theta)}\end{aligned}$$

$$\begin{aligned}\Pr(AB|AB) &= k_2 + k_1\frac{2\theta + (1-\theta)(p_A + p_B)}{2(1+\theta)} \\ &\quad + k_0\frac{2[\theta + (1-\theta)p_A][\theta + (1-\theta)p_B]}{(1+\theta)(1+2\theta)}\end{aligned}$$

Paternity Index with Homozygous Mother

G_M : Mother's genotype; G_C : Child's genotype;
A_M : Maternal allele; A_P : Paternal allele;
G_{AF} : Alleged father's genotype; PI: Paternity index.

G_M	G_C	A_M	A_P	G_{AF}	PI
AA	AA	A	A	AA	$\frac{1+3\theta}{4\theta+(1-\theta)p_A}$
				AB	$\frac{1+3\theta}{2[3\theta+(1-\theta)p_A]}$
	AB	A	B	BB	$\frac{1+3\theta}{2\theta+(1-\theta)p_B}$
				AB	$\frac{1+3\theta}{2[\theta+(1-\theta)p_B}$
				BC	$\frac{1+3\theta}{2[\theta+(1-\theta)p_B]}$

Paternity Index with Heterozygous Mother

G_M : Mother's genotype; G_C : Child's genotype;
A_M : Maternal allele; A_P : Paternal allele;
G_{AF} : Alleged father's genotype; PI: Paternity index.

G_M	G_C	A_M	A_P	G_{AF}	PI
AB	AA	A	A	AA	$\frac{1+3\theta}{3\theta+(1-\theta)p_A}$
				AB	$\frac{1+3\theta}{2[2\theta+(1-\theta)p_A]}$
	AC	A	C	CC	$\frac{1+3\theta}{2\theta+(1-\theta)p_C}$
				AC	$\frac{1+3\theta}{2[\theta+(1-\theta)p_C]}$
				CD	$\frac{1+3\theta}{2[\theta+(1-\theta)p_C]}$

GENETIC GENEALOGY

Identity by Descent

Two alleles from the same ancestral allele are identical by descent.

Individuals that share alleles identical by descent are related.

Individuals may share 0,1 or 2 pairs of alleles identical by descent: e.g. they may have both, either or neither of their maternal and paternal alleles identical by descent. The probabilities of these three states are k_0, k_1, k_2.

The kinship coefficient of two people is $\theta = k_2/2 + k_1/4$.

STR Kinship Coefficients

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	$\frac{1}{4}$	$\frac{3}{4}$	$\frac{1}{16}$
nth cousins	0	$\left(\frac{1}{4}\right)^n$	$1 - \left(\frac{1}{4}\right)^n$	$\left(\frac{1}{4}\right)^{n+1}$
Unrelated	0	0	1	0

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

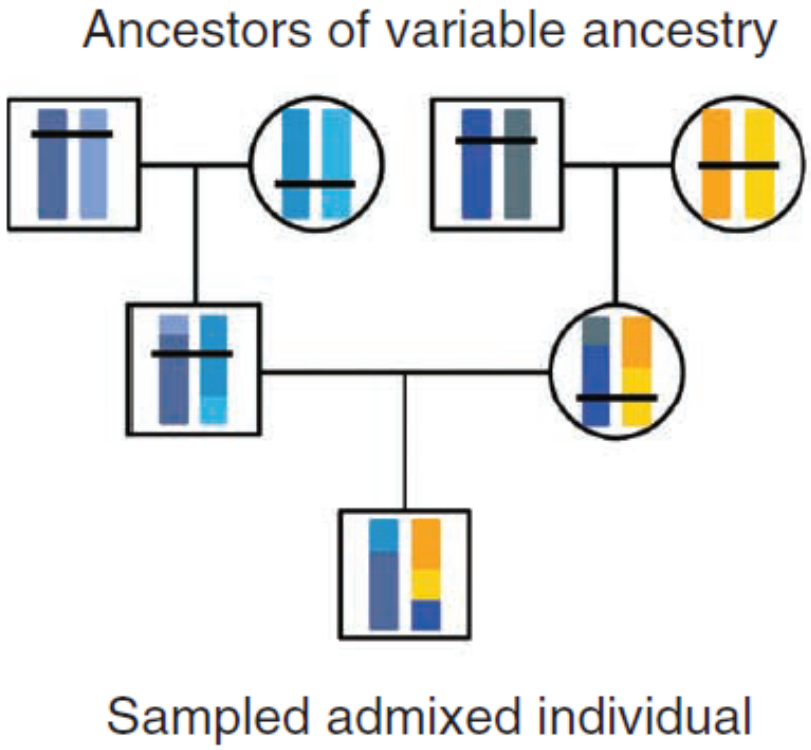
STR Kinship Coefficients

These kinship coefficients with forensic STR panels are not good for distinguishing different types of relatives beyond half sibs. Difficult even to separate half sibs from full sibs.

SNP panels, with up to a million SNPs allow distinguishing even distant cousins. A different statistical measure is used, that takes (lack of) recombination into account.

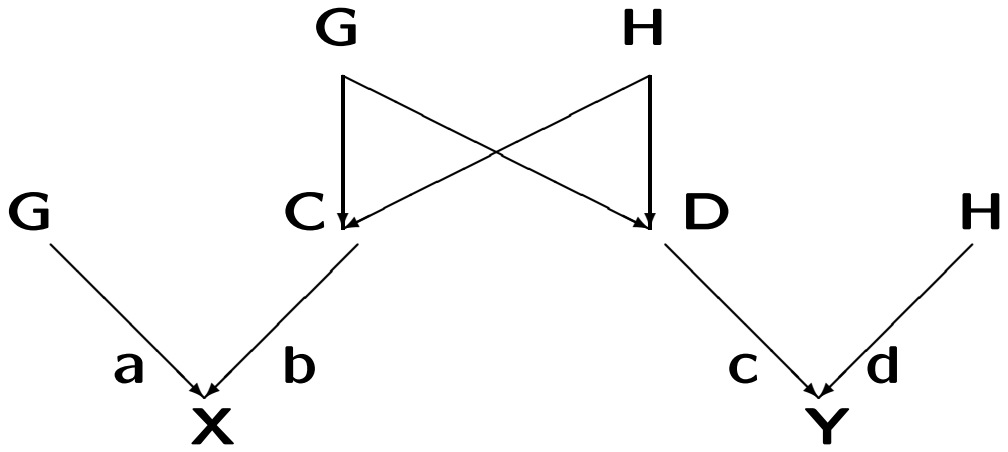
Recombination

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

First Cousins



X, Y are first cousins, and are expected to share identical alleles from one grandparent with probability 1/16.

But most parts of their genomes will not share identical alleles and some blocks will have identity across the block.

The Shared cM Project

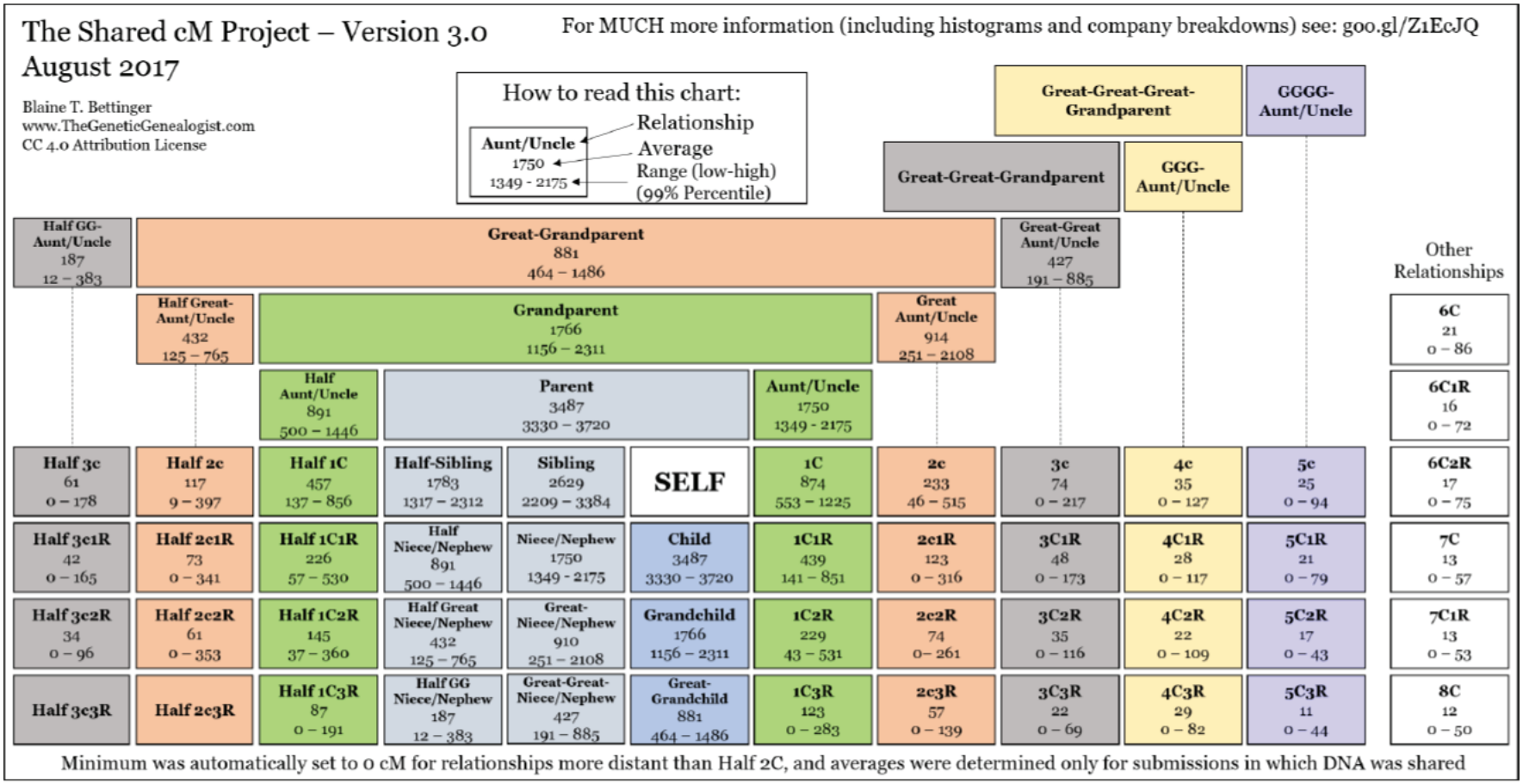
https://thegeneticgenealogist.com/

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

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

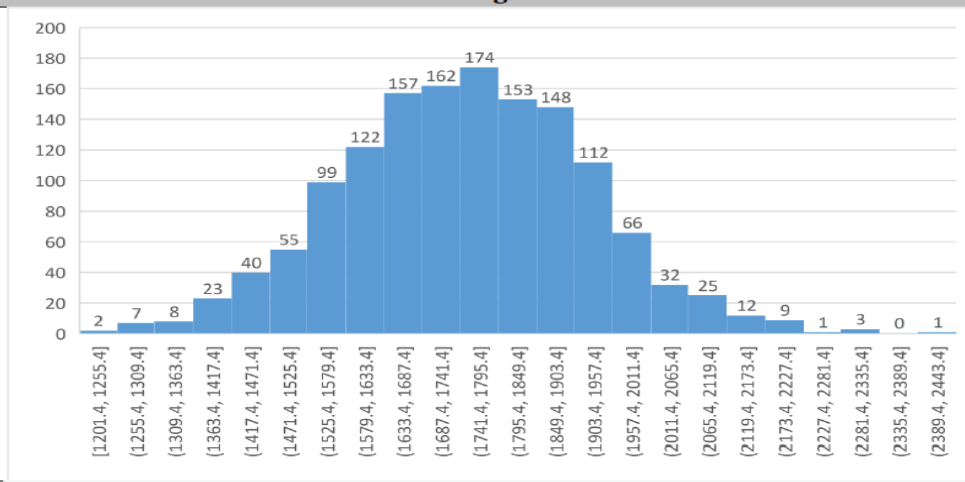
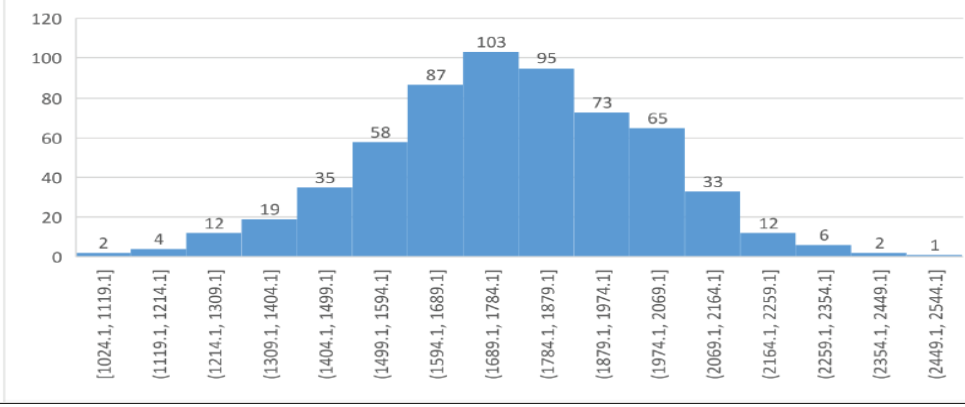
Blaine T. Bettinger
www.TheGeneticGenealogist.com

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	

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

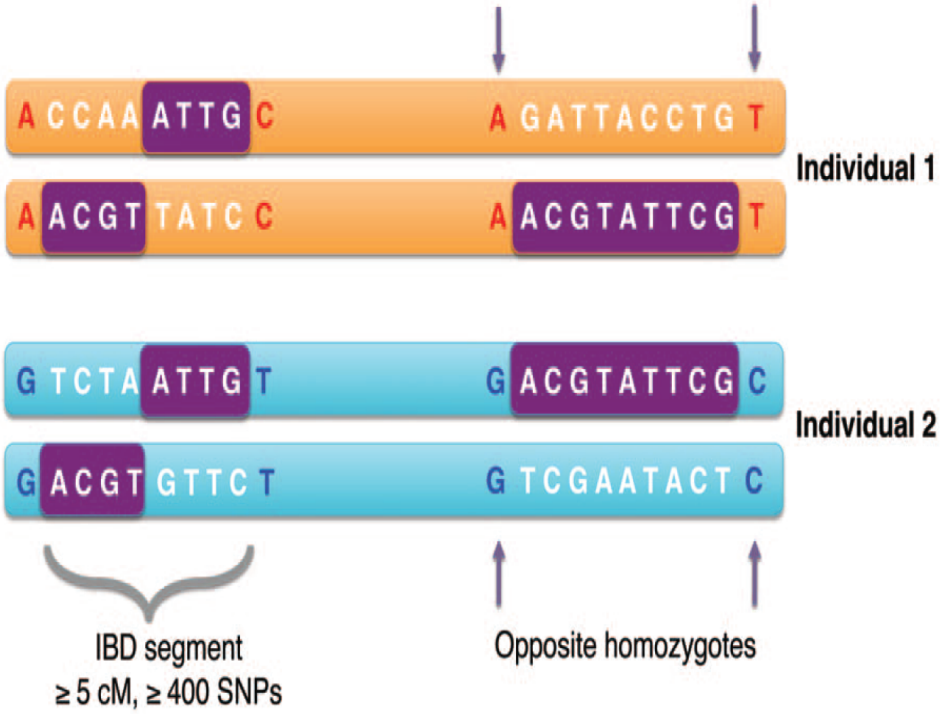


Figure 1. Schematic of IBD_{half} inference method. IBD_{half} segments were inferred from unphased genotype data where a series of alleles were identical by state for *at least one* of the homologous chromosomes in a given pair of individuals. IBD segments are indicated in purple. The boundaries of 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

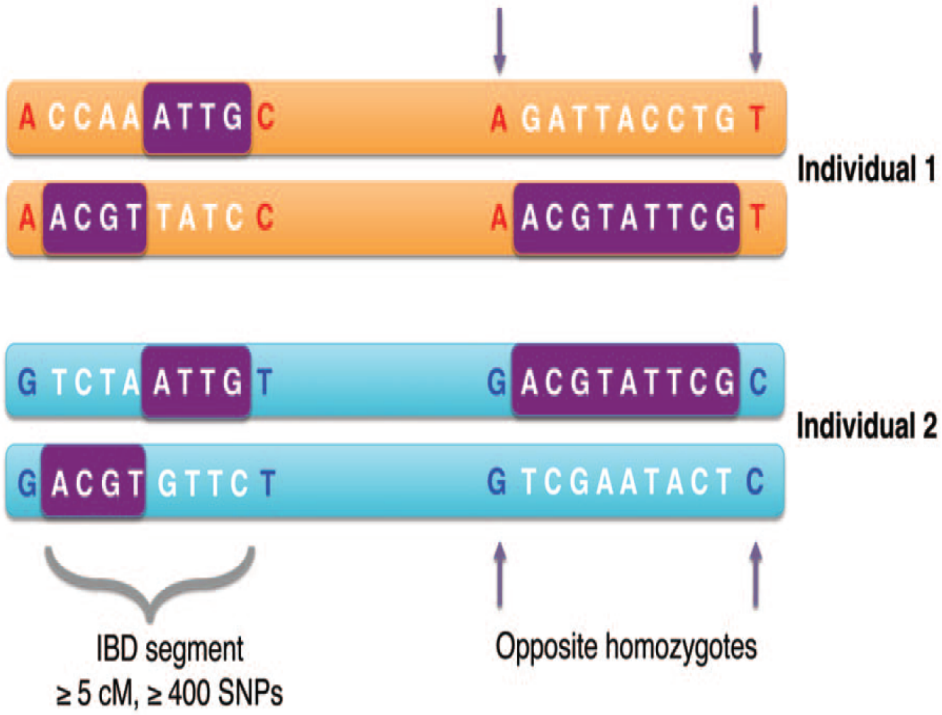


Figure 1. Schematic of IBD_{half} inference method. IBD_{half} segments were inferred from unphased genotype data where a series of alleles were identical by state for *at least one* of the homologous chromosomes in a given pair of individuals. IBD segments are indicated in purple. The boundaries of 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 $C1, C2$.

E and $C1$ have two of their four grandparents in common. Think of the four grandparents of $C1$ and trace their descendants $D1$: these 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 $D2$: these are the parents, uncles, aunts and cousins of $C2$.

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

CEU Example

A CEU individual in the 1000Genomes project appears to have parents who were first cousins. Using 1,000 windows of 1000 SNPs, chromosome 22 shows: