www.sciencemag.org/cgi/content/full/science.aau4832/DC1

Supplementary Materials for

Identity inference of genomic data using long-range familial searches

Yaniv Erlich*, Tal Shor, Itsik Pe'er, Shai Carmi

*Corresponding author. Email: erlichya@gmail.com

Published 11 October 2018 on *Science* First Release
DOI: 10.1126/science.aau4832

This PDF file includes:

Materials and Methods

1. Measure shared IBD with the MyHeritage database

The MyHeritage database mainly consists of individuals that were tested with the MyHeritage DNA product. Briefly, individuals swab the inner side of their cheeks using a sterile absorbent tipped applicator (HydraFlock). After sampling DNA on the inner side of the cheeks, the participant places the tip of the applicator in a vial that is filled with a standard lysis buffer. The DNA is transferred to a CLIA certified lab, where is genotyped with an Illumina OmniExpress genome-wide genotyping array that contains 700,000 SNPs. Another route for participants to enter the database is by uploading their raw genotype files from other DTC companies. Currently, the website supports uploads from 23andMe (v1-v4 kits), Ancestry (all versions), and FTDNA (all versions). All participants have agreed to the MyHeritage's Terms and Conditions that permits genetic analysis of their data.

To measure the probability of finding a relative above a certain shared IBD, we took the results from the standard DNA processing pipeline of MyHeritage, which lists all IBD segments above 6cM for pairs of individuals. For this study, we used 1,277,872 samples. IBD segments for these samples were stored in a special research database in a de-identified format capable of fast computing.

Next, we queried the database with various levels of minimal shared cM between the relatives. In our experience, customers tend to purchase more than one kit and hand the other kit to a close family member. To mitigate ascertainment bias, we deliberately excluded all pairs of individuals with total IBD length above 700cM, who are likely to be first cousins or closer relatives. As the service offers individuals to document their family trees, we also excluded pairs of relatives with a known genealogical path whose kinship coefficient is 0.125 or higher, such as first cousin, grandparents, uncles, parents, and siblings. We then queried the database with thresholds for total shared IBD growing from 30cM to 600cM and counted the number of individuals with at least one match.

To calculate the genetic ethnicity of each user, we used the standard results of the MyHeritage ethnicity pipeline. This pipeline reports 42 possible ethnicities based on a reference dataset of over 5000 samples collected from MH participants that consented for this process and presented homogenous ethnicity as reported by the place of births of their ancestors. For the purpose of this study, we assigned each ethnicity to one out of nine groups that were mainly based on subcontinental regions (**fig. S2; Table S3**).

2. Measure shared IBD with the GEDmatch database.

GEDmatch employs a unique model where the report of the genetic matches of any GEDmatch user who opted-in to the "One-to-many DNA comparison" is publicly available. The website allows anyone with an Internet connection to view the list of matching relatives sorted by the total shared IBD segments. The list includes key details about each match, such as contact information of each relative, summary statistics of the IBD segments, and in some cases also pedigree information (but not the raw DNA data).

We randomly ascertained 30 existing GEDmatch users. Importantly, we did not transfer any genetic information from MyHeritage to this website. Rather, we simply observed existing GEDmatch users, whose matching results are publicly available.

To avoid ascertainment biases, we used a random number generator to draw potential GEDmatch profile ID numbers. Each profile ID refers to a user and has the convention xdddddd, where x is a letter that signifies the DTC company that generated the uploaded genotype file (e.g. A: ancestry, H: MyHeritage, M: 23andMe, T: FamilyTreeDNA) and d is a digit between 0 to 9. After generating a profile ID number, we manually inserted the GEDmatch profile ID number to the search box in the "One-to-many DNA comparison" tool.

We selected the default parameters that requires IBD matches to be at least 7cM long. For each match list, we excluded all matches above 700cM, similar to the exclusion process with the MyHeritage data described above. Then, we selected the top match. Finally, we examined the list of the top match for the 30 GEDmatch users and filtered the list according to various cM thresholds, as reported in the main text. The 95% confidence intervals were produced using the Wilson Score Interval.

3. Population genetic theoretical evaluation of long range familial search

The problem

Consider a database of genotyped individuals from a defined population and the DNA of a person of interest, called the target. We would like to identify the target by finding his/her relatives in the database. We calculate below the probability to find one or more relatives given the population size, the database size, and the matching parameters.

The model's assumptions

1. We assume a monogamous Wright-Fisher model, similar to that of Shchur and Nielsen (*25*). In the current generation, the population has N males and N females, organized in N couples. Each individual in the current generation chooses its parents (i.e., a couple) randomly out of all couples in the previous generation.

2. The number of children per couple is $r > 2$. Thus, the population size (the number of couples) at generation g before the present is $N(g) = N(r/2)^{-g}$.
3. Individuals are diploid and we consider only the autosomal genome.
4. The database has R individuals.
5. The genome of the target individual is compared to those of all individuals in the database, and identical-by-descent (IBD) segments are identified. We assume that detectable segments must be of length $\geq m$ (in Morgans). We further assume that in order to confidently detect the relationship (a "match"), we must observe at least s such segments.
6. We only consider relationships for which the common ancestors have lived $g \leq g_{\text{max}}$ generations ago. For example, $g = 1$ for siblings, $g = 2$ for first cousins, etc. All cousins/siblings are full.
7. The number of matches between the target and the individuals in the database is counted. If we have more than t matches, we declare that there is sufficient information to trace back the target. Typically, we simply assume $t = 1$.

Derivation

The probability of a sharing a pair of ancestors

Consider two individuals: the target and a single individual in the database. g generations before the present, each one of them has 2^{g-1} ancestral couples. For example, each individual has one pair of parents ($g = 1$), two pairs of grandparents ($g = 2$), four great-grandparents ($g = 3$), and so on. Under the assumption that $2^g \ll N(g)$, the probability that the two individuals have one ancestral couple is approximately (*26*):

$$(1)\quad P(\text{shared ancestral couple } g \text{ generation ago}) \approx \frac{2^{g-1}2^{g-1}}{N(g)} = \frac{2^{2g-2}}{N(g)}.$$

We ignore the possibility of sharing more than one ancestral couple, assuming that $2^{2g} \ll N(g)$.

The probability to share an ancestral couple for the first time at generation g is approximately:

$$(2)\quad P(\text{first sharing mating pair at } g) \approx \frac{2^{2g-2}}{N(g)} \prod_{g'=1}^{g-1} \left(1 - \frac{2^{2g'-2}}{N(g')}\right).$$

The probability of a match given a shared mating pair

Next, we determine the probability that a target and an individual in the database are identified as genetic relatives using their IBD segments, conditioned that they share an ancestral couple g generations ago. This probability boils down to the probability that the target and the relative share at least s IBD segments longer than m.

We use a simple approximation that the genome can be broken into blocks that are inherited independently. If the ancestors have lived g generations ago, the two individuals are separated by $2g$ meioses. Given that the total genome length is roughly $L = 35$ Morgan, there are thus on average $2gL \approx 70g$ recombination events between the two individuals. Since blocks are bounded by either recombination or chromosome ends, a rough approximation for the number of blocks is $2Lg + 22$, as in ref. (*27*).

Next, we calculate the probability that a genomic block of a pair of genealogical relatives is identical by descent. For simplicity, let's name them Alice and Bob and assume that they are connected through their maternal sides. As we stated above, Alice has 2^{g-1} ancestral couples and therefore $\frac{2^{g-1}}{2} = 2^{g-2}$ ancestral couples from her maternal side. Similarly, Bob also has 2^{g-2} ancestral couples from his maternal side. Since the pair only share one ancestral couple, the probability that they pick the shared ancestral couple for a genomic block is $\frac{1}{2^{g-2}} \cdot \frac{1}{2^{g-2}} = 2^{4-2g}$. The shared ancestral couple has four chromosomes. Therefore, Alice and Bob have $1/4$ chance to pick the same chromosome. Thus, in overall, the probability to share a block as identical-by-descent is $2^{4-2g} \cdot \frac{1}{4} = 2^{2-2g}$. For example, first cousins have a shared grandparental couple $(g = 2)$. Therefore, they have $2^{2-4} = 25\%$ chance that a diploid genomic block contains an identical by descent segment.

Next, we determine the probability of the IBD segment to be over m Morgans in order to be detected, given that the pair share a specific block. The length of the segment is exponentially distributed with a mean of $1/(2g)$ Morgans. Thus, if x is the segment length, $P(x) = 2ge^{-2gx}$. The probability of the segment length to exceed m is $\int_m^\infty 2ge^{-2gx}dx = e^{-2mg}$. Thus, in each block, the probability of sharing a detectable IBD segment is:

$$(3) \quad P(\text{IBD}) = \frac{e^{-2mg}}{2^{2g-2}}.$$

Assuming that blocks are independent, the probability to share k blocks is binomial: $P(\text{share } k \text{ blocks}) \sim \text{Bin}(k; n, p)$, with $n = 2Lg + 22$ and $p = P(\text{IBD})$ above. To declare a match, we need at least s segments of at least length m. Thus, given a shared mating pair g generations ago, the probability to observe a match is

$$(4) \quad P(\text{match}|g) = 1 - \sum_{k=0}^{s-1} \text{Bin}\left(k; 2Lg + 22, \frac{e^{-2mg}}{2^{2g-2}}\right).$$

The number of matches to the database

The probability of declaring a match between the target and a random individual in the database is simply the sum of the product of Eqs. (2) and (4) over all g,

$$(5) \quad P(\text{match}) = \sum_{g=1}^{g_{\max}} \left[\prod_{g'=1}^{g-1} \left(1 - \frac{2^{2g'-2}}{N(g')}\right) \frac{2^{2g-2}}{N(g)} P(\text{match}|g)\right].$$

To calculate the expected number of matches between the target and the entire database, we assume that the probability of a match to each individual in the database is independent. This approximation follows a result of Shchur and Nielsen (*25*) (Eq. (7) therein), who showed that for a large population, the probability of an individual to have a relative in the database is as if the database individuals were independent. Under this assumption, the number of matches is binomial, with $n = R$ and $p = P(\text{match})$ from Eq. (5). To identify an individual, we need to find at least t matches in the database. Thus,

$$(6)\quad P(\text{identify}) = 1 - \sum_{k=0}^{t-1} \text{Bin}\big(k; R, P(\text{match})\big)$$

The probability of re-identification through both parents

We can also consider a more involved scenario in which triangulation necessitates detecting relatives from both sides of the family of the person of interest, for example finding 1C1R from the mother side and 2C from the father side.

Given a match to a database individual, the path to the shared ancestors goes through the target's father or mother with equal probability. Given k matches, the probability that all matches go through the mother is $\left(\frac{1}{2}\right)^k$ and the same for the father. The probability of re-identification through both sides can thus be written as

$$(7)\quad P(\text{identify, both sides}) = \sum_{k=t}^{R} \left(1 - \left(\frac{1}{2}\right)^{k-1}\right) \text{Bin}\big(k; R, P(\text{match})\big)$$

Matches between cousins once removed

Our derivation is straightforward to generalize to cases when the target and database pairs of relatives belong to different generations, such as second cousins one removed. For example, suppose that the common ancestor has lived $g + 1$ generations from one individual and g generations from the other individual; namely, the two individuals are g-generations relatives once removed. We denote by $N(g)$ the population size at g generations ago with respect to the younger individual. The probability to share a mating pair at generation $g \geq 1$ from the **older** individual is now $\approx \frac{2^{2g-1}}{N(g+1)}$ (as opposed to $2^{2g-2}/N(g)$ when ages were equal). The probability of a match given a shared mating pair is the same as in the above derivation, except that the number of meiosis is now $2g + 1$ (instead of $2g$), and the probability for the two individuals to inherit the same chromosome from the common ancestor is $1/2^{2g-1}$ (instead of $1/2^{2g-2}$). The remaining derivation is identical to that above.

We found numerically that the probability of identification when considering “once removed” matches up to $g_{\max}$ generations ago is intermediate between the probabilities obtained using $g_{\max}$ and $g_{\max}+1$ in the main derivation (Eq. (6)), usually closer to $g_{\max}+1$. To be on the conservative side, when reporting results for once-removed relationships, we sum over all generations up to $g_{\max}-1$ only.

In reality, databases may have matches for both regular cousins and once-removed cousins. To incorporate both types of relatives into the probability of identification, we made two simplifying assumptions. First, we assumed that the target belongs the current (most recent) generation. Second, we assumed that the database consists of individuals from both the current and previous generations, with proportions equal to their proportions in the total population (i.e., $1\colon (r/2)$, given a population growth rate of r offspring per couple per generation). Thus, given that the total database size is R, the number of individuals it contains from the current and previous generations are $R\frac{r/2}{1+r/2}$ and $R\frac{1}{1+r/2}$, respectively. For a given set of matching parameters, we denote by $P_{0r}(\text{match})$ and $P_{1r}(\text{match})$ the probability of detecting a match with a cousin or a cousin once removed, respectively. Given that we still require t matches for identification, the probability of identification becomes

$$(8)\quad P(\text{identify}) = 1 - \sum_{k=0}^{t-1}\sum_{k_0=0}^{k} \text{Bin}\left(k_0; \frac{Rr/2}{1+r/2}, P_{0r}(\text{match})\right)\cdot \text{Bin}\left(k-k_0; \frac{R}{1+r/2}, P_{1r}(\text{match})\right).$$

Eq. (8) was used for plotting Figure 1B. The relevant R code is provided below.

R code

```
genome_size = 35
num_chrs = 22

p_match = function(g,m,min_num_seg)
{
  m = m/100
  f = exp(-2*g*m)/2^(2*g-2)
  pr = 1 - pbinom(min_num_seg-1,num_chrs+genome_size*2*g,f)
  return(pr)
}

p_match_or = function(g,m,min_num_seg)
{
  m = m/100
  f = exp(-(2*g+1)*m)/2^(2*g-1)
  pr = 1 - pbinom(min_num_seg-1,num_chrs+genome_size*(2*g+1),f)
  return(pr)
}

coverage = function(Ks,maxg,N_pop,r,m,min_num_seg,min_num_rel, rep_direct =
rep(1,10), rep_or = rep(1,10))
{
  N = N_pop/2 # convert pop size to couple size
  pr_succ = length(Ks)
  for (i in 1:length(Ks))
  {
    K = Ks[i]
    K_same = round(K * (r/2) / (1+r/2))
    K_or = round(K * 1 / (1+r/2))

    p_no_coal = numeric(maxg)
    p_coal = numeric(maxg)
    # OR: Once Removed
    p_no_coal_or = numeric(maxg)
    p_coal_or = numeric(maxg)
    Ns = N*(r/2)^(-(1:(maxg+1)))
    tot_p = 0
    tot_p_or = 0
    for (g in 1:maxg)
    {
      f = 2^(2*g-2)/Ns[g]
      f_or = 2^(2*g-1)/Ns[g+1]
      if (g>1) {
        p_coal[g] = p_no_coal[g-1] * f
        p_no_coal[g] = p_no_coal[g-1] * (1-f)
        p_coal_or[g] = p_no_coal_or[g-1] * f_or
        p_no_coal_or[g] = p_no_coal_or[g-1] * (1-f_or)
      } else {
        p_coal[g] = f
        p_no_coal[g] = 1-f
        p_coal_or[g] = f_or
        p_no_coal_or[g] = 1-f_or
      }

      tot_p = tot_p + p_coal[g] * p_match(g,m,min_num_seg) * rep_direct[g]
      if (g<maxg) {
        tot_p_or = tot_p_or + p_coal_or[g] * p_match_or(g,m,min_num_seg) *
rep_or[g]
      }
    }

    pr_no_succ = 0
    for (n in 0:(min_num_rel-1))
    {
      for (n_or in 0:n)
      {
        pr_no_succ = pr_no_succ + dbinom(n_or,K_or,tot_p_or)*dbinom(n-
n_or,K_same,tot_p)
      }
    }
    pr_succ[i] = 1 - pr_no_succ
  }
  return(pr_succ)
}


# Ks: A vector of database sizes
# maxg: Maximum relatedness to consider (1: sibs, 2: 1st cousins, 3:
2nd cousins...)
# N: Population size
# r: Mean number of children per mating pair (=per family), so 2 for a
constant size population, >2 for expanding population, <1 for
contracting population
# m: Maximum length in cM of a detectable segment
# min_num_seg: Minimum number of segments to declare a match
# min_num_rel: Minimum number of detected matches (=relatives)
# to declare success of identification
```

To produce **Fig. 1B**, we used the following parameters with the R code above:

```
N = 250000000 #population size
num_K = 10000 #number of data points between 0 to 1
m = 6 #minimal cM
min_num_seg = 2 #number of segments
r = 2.5 #number of kids per couple
Ks = round(seq(from=N/num_K,to=N,length.out=num_K))

c1 = coverage(Ks,maxg=2,N_pop,r,m,min_num_seg=2,min_num_rel=1)
c2 = coverage(Ks,maxg=3,N_pop,r,m,min_num_seg=2,min_num_rel=1)
c3 = coverage(Ks,maxg=4,N_pop,r,m,min_num_seg=2,min_num_rel=1)
c4 = coverage(Ks,maxg=5,N_pop,r,m,min_num_seg=2,min_num_rel=1)
```

Detecting matches by the total length of shared IBD segments

One of the model's assumptions is that matches are detected whenever the number of shared IBD segments exceeds a cutoff. However, the adversary may consider only matches above a certain total IBD length in cM. Our model can be revised accordingly to evaluate that probability to detect relationships with IBD above a certain cM threshold, as follows.

In Eq. (3), we found that given a genealogical shared ancestor g generations ago, the probability of the target and database individuals to share an IBD segment at each locus is $P(\text{IBD}) = \frac{e^{-2mg}}{2^{2g-2}}$. The total number of shared segments, which we denote as n_s, can be approximated as Poisson with mean $(2Lg + 22) \cdot \frac{e^{-2mg}}{2^{2g-2}}$. The total length (in Morgan) of the shared segments, which we denote as ℓ_T, is the sum of n_s segment lengths, where each segment length is an exponential random variable with rate $2g$ conditioned to be longer than the minimum length m. The distribution of ℓ_T is

$$P(\ell_T) = \sum_{n_s=1}^{\infty} \text{Pois}\left(n_s; (2Lg + 22) \cdot \frac{e^{-2mg}}{2^{2g-2}}\right) \cdot P(\ell_T \mid n_s). \tag{9}$$

Given the number of segments n_s, the total length covered by those segments has an Erlang distribution with shape parameter n_s and rate $2g$, except that distribution is a function of $\ell_T - n_s m$ rather than ℓ_T (i.e., it is shifted by $n_s m$, which is the minimal total length). Given a cutoff ℓ_c for detection of the match, the probability of a match is $P(\ell_T > \ell_c)$. This can be written using the upper incomplete gamma function, $\Gamma(a,x) = \int_x^\infty y^{a-1}e^{-y}dy$.

$$(10)\ P(\text{match}) = P(\ell_T > \ell_c) = \sum_{n_s=0}^{\infty} \text{Pois}\left(n_s; (2Lg + 22) \cdot \frac{e^{-2mg}}{2^{2g-2}}\right) \cdot \frac{\Gamma(n_s, 2g(\ell_c - n_s m))}{\Gamma(n_s)}$$

To compute the sum in Eq. (10), it is sufficient to sum over the first few terms for which the Poisson distribution is non-negligible. The probability of identification is the same as above (Eq. (6)), and the extension to cousins once removed is also straightforward (changing $2g$ to $2g + 1$).

In practice, to examine the effect of a total cM cutoff on the probability of identification, we took an empirical approach based on reports from the shared cM project (*16*). Specifically, for a given genealogical distance g, we used the empirical reported range of total shared cM to compute the probability of the total cM exceeding a cutoff. We multiplied that empirical probability of detection by the above calculated probability of detecting a match based on a minimum number of segments (Eq. (4)). The rationale behind this approach is that in practice, background (population-level) IBD sharing may change the total IBD length shared. Thus, while we may still require a minimum of (say) two particularly long segments to detect a relationship, we would also require the total length shared to exceed a cutoff based on empirical data from validated relationships.

Specifically, to produce **fig. S5**, we ran the following commands:

```
c5_50cM = coverage(Ks,maxg=6,N,r,m,min_num_seg=2,min_num_rel=1, rep_direct =
p50cM_cousins, rep_or = p50cM_1R)
c5_100cM = coverage(Ks,maxg=6,N,r,m,min_num_seg=2,min_num_rel=1, rep_direct =
p100cM_cousins, rep_or = p100cM_1R
c5_200cM = coverage(Ks,maxg=6,N,r,m,min_num_seg=2,min_num_rel=1, rep_direct =
p200cM_cousins, rep_or = p200cM_1R)
c5_300cM = coverage(Ks,maxg=6,N,r,m,min_num_seg=2,min_num_rel=1, rep_direct =
p300cM_cousins, rep_or = p300cM_1R)
c5_400cM = coverage(Ks,maxg=6,N,r,m,min_num_seg=2,min_num_rel=1, rep_direct =
p400cM_cousins, rep_or = p400cM_1R)
c5_500cM = coverage(Ks,maxg=6,N,r,m,min_num_seg=2,min_num_rel=1, rep_direct =
p500cM_cousins, rep_or = p500cM_1R)
```

The p50cM_cousins and p50cM_1R variables (and other variables with different cM cutoffs) are vectors that represent the fraction of Xth-cousins and Xth-cousins once removed, respectively, to be above a certain IBD threshold. For example, for the probabilities of passing 50cM, we used:

```
p50cM_cousins = c(0, #bro, set to zero because we ignore these relationships
                  0, #1C, similarly set to zero
                  (1590-16)/1590, #2C. The numbers are taken from the histogram of the
shared cM Projects
                  (1791-230+183+120)/1791, #3C
                  (75/2+49+39+24+18+14+6+2+2+1+2+2)/998, #4C
                  (15+7+7+7+1+3)/429) #5C

p50cM_1R = c(0, #uncle
             1, #1C1R
             (2064-(162*0.3+97+27))/2064, #2C1R
             (1736-199-221-231-253-198)/1738, #3C1R
             (934-45-67-105-154-178-199)/934) #4C1R
```

The numbers in these vectors were taken from the shared cM project (*16*), besides for siblings, uncles, and 1C that were set to zero because we have filtered these relationships from our empirical results.

A comparison to previous models

During the preparation of this manuscript, Michael Edge and Graham Coop (henceforth EC) have posted a similar analysis in a blog article that computes the probability of detecting a match in a database (*28*). Our model is generally similar to the EC model with several key differences. First, EC have modeled the historical population size based on the census size of the United States and additional assumptions, whereas we assumed a certain size for the current generation, and an exponential contraction at a constant rate going backwards. Second, we considered IBD segments that are longer than a length threshold m, which is common in relative matching pipelines, whereas the EC model considered segments of all lengths. Third, we expanded the theoretical model in several directions. Specifically, we computed the probability of a match (a) to cousins once-removed (also incorporating multiple generations into the database); (b) to both parents of the target; and (c) under the assumption of a total IBD length cutoff rather than a minimal number of segments (using both theoretical and empirical approaches). Fourth, in our derivation, we first modeled the probability of a match (at any generation) between the target and a single database individual. Using that probability, we then computed the probability of a minimum number of matches between the target and the entire database (i.e., the probability of identification). In contrast, the EC model computed the mean number of database matches that corresponds to each generation. Other modeling differences (e.g., using a binomial vs Poisson distributions) also exist but have smaller effects.

4. Pruning the search space using demographic identifiers

In this section, we attempt to evaluate the amount of effort required to identify an individual based on a detected DNA match. To identify the target individual, we would need to examine all possible relatives of the match that are consistent with the genetic distance between the match and the target. Our task is to measure the search space, namely, the number of such relatives, and to assess the power of demographic identifiers to reduce the search space.

We focused on the case when the match shares with the target IBD segments of total length of ~100cM, which is at the lower end of the relatedness level that we still consider useful for identification. Such matches correspond to the following types of relatives: 1C2R, 2C1R, 3C, 2C, and 2C2R. We did not take into account rarer types of relatives with a similar genetic distance, such as half-second cousins, as these are less likely to be encountered. However, we do distinguish the direction of the removal of the cousins. For example, if Bob is Alice's 1st cousin once removed, Bob can be (a) the 1st cousin of Alice's parent; or (b) the son of Alice's 1st cousin. To distinguish between these two possibilities, we will refer to cousin relationships with non-zero removals as "up" and "down". For example, if Bob is Alice's mom 1st cousin, we shall say that Bob is Alice's 1st cousin once-removed up (1C1R-u) and that Alice is Bob's 1st cousin once-removed down (1C1R-d). Genetically, the "up" and "down" relationships have the same IBD characteristics. But from a demographic perspective, the up and down relationships are quite different. For example, if Bob is Alice's 1st cousin once removed up, we expect him to be older than Alice and vice versa.

In the first step, we computed the total number of relatives of types 1C2R, 2C1R, 3C, 2C, and 2C2R, including all possible up and down relationships (e.g. 1C2R-u and 1C2R-d). Assuming that each couple gives birth to 2.5 children on average, and that all children reach a fertility age (similar to ref. (*17*)), the average number of xC relatives is $f(x) = 2^x \cdot 1.5 \cdot 2.5^x$, e.g., 7.5 for 1C, 37.5 for 2C, and 187.5 for 3C. The number of xCyR-d relatives is $2.5^y f(x)$, e.g., 46.875 for 1C2R-d and 93.75 for 2C1R-d. The number of xCyR-u relatives is $2.5^{-y} f(x+y)$, e.g., 75 for 2C1R-u and 30 for 1C2R-u.

In total, we estimate that there are on average 855 relatives that could match a genetic distance of 100cM, who all need to be examined to identify the target. In the next paragraphs, we use a combination of simulations and real data from our previous study of population-scale family trees (*19*) to evaluate the power of demographic characteristics to reduce the number of relatives to follow up. The family trees dataset has been subject to extensive types of validation, including accuracy assessment using genetic data and concordance analysis with government-based demographic data.

Geography

We used our genealogical records to analyze the geographic distance between relatives. We analyzed 145,658 pairs of relatives encompassing 1C2R, 2C1R-up/down, 3C, 2C, and 2C2R-up/down (**Table S4**), considering only pairs where at least one individual was born in the US between 1940 to 2010. We then calculated the geographical distance based on the longitude and latitude of the birth locations.

We assumed that the geographic location of the target is known to within 100 miles, and considered a conservative scenario in which the residence of the matched relative coincides with the target; otherwise, the number of relatives of the match that are within the search radius will be even lower. Thus, we need to consider only relatives of the match living within 100 miles of the match. We found that less than 30% of all 1st cousins once removed, and 51% of all 2nd cousins, are expected to be present within the search space centered around the match.

Similarly considering the other types of relatives, we find that, on average, only 369 relatives of the above types live within 100 miles from the match.

We also investigated whether there is a difference in the power of geographic information to implicate an individual between highly populated areas versus less populated areas. We repeated the analysis after grouping the cousin pairs into three categories: (A) pairs where at least one person was born in one of the top 10 most populous cities in the US; (B) pairs where at least one person was born in a place that was mentioned only up to two times in the dataset (about 400 times less than New York pairs). In this group, we had places like Eureka, CA, that has approximately 25,000 residents and Plant City, FL, that has approximately 35,000 residents. We did not find any major differences between cases where the simulated suspect lives in highly populated areas versus less populated areas. For the highly populated areas (group A), a 100 miles radius retained 56% of the relatives compared to 55% of the less populated areas (group B).

Age

To analyze the age dispersion of pairs of relatives, we conducted extensive simulations that were further validated with a large set of 3rd cousins.

Simulations: Consider a pair of cousins (of any type) named Alice and Bob. These cousins necessarily descend from a pair of siblings, which we denote as Anna (the ancestor of Alice) and Brad (the ancestor of Bob). In the following, we describe a method to simulate the age difference between the cousins. Genealogically, the age differences can be expressed using three processes: (i) the difference between the ages of Anna and Brad (i.e., the age of Anna minus the age of Brad), denoted by $s;$ (ii) the sum of the parental age at birth of i consecutive descendants of Anna, denoted by v_i (iii) the sum of the parental age at birth of j consecutive descendants of Brad, denoted by u_j. For example, if Bob is the 2nd cousin once-removed up (2C1R-u) of Alice, the difference (in absolute value) between the ages of Alice and Bob is $v_3 - u_2 - s$. In general, if Bob is Alice's x-cousin y-removed up, the difference in the year of birth is $v_{x+y} - u_x - s$, while if Bob is Alice's x-cousin y-removed down, the difference in the year of birth is $u_{x+y} - v_x + s$.

To simulate v, u, and s, we examined the distribution of parental age at birth using 1,752,000 parent-offspring pairs that reflect the highest quality of our data with exact date of births and birth places. These pairs were born between 1650 to 1950. To reduce the chance of errors, we retained only pairs where the parental age at birth was between 10 to 60 years, which excluded 0.06% of the pairs, leaving 1,741,000 parent-offspring pairs for subsequent analysis. The average parental age at birth was 31.7 years, close to previous genealogical studies of generation times in the Western world (*29*).

We used a similar process to create a histogram of the age difference between full siblings using pairs that were born between 1650 to 1910. We retained only pairs of siblings with an age difference of up to 50 years, leaving 879,000 pairs of siblings for subsequent analysis. To simulate an instance of s, we sampled an event from the probability mass function of the sibling age difference. To simulate an instance of v_x or u_x, we randomly sampled x events according to the probability mass function of the parental age at birth and summed them together.

We simulated 100,000 age differences of each type of relative. We then calculated the age distribution of each class, and mixed the distributions according to the number of individuals in each relationship class who are expected to live less than 100 miles from the match. The entropy of the distribution with 10yr bins was 3.95bits and the entropy of the 1yr bins was 7.26bits.

Given that we know the age of the target within 10yr or 1yr interval, we will have to follow up only on those relatives of the match at the same age group as the target. Thus, in the most conservative scenario, the age of the target would fall under the tallest bin of the histogram. We measured the size of this bin for 10yr and 1yr intervals and reported the results in the main text.

As another layer of validation, we repeated the analysis but this time we only utilized pairs of parent-child that were born after 1850 and sibling pairs where at least one sibling was born after 1850. We hypothesized that these pairs can be more relevant to modern families. The number of parent-child pairs was 966,000 and the number of sibling pairs was 336,000. Despite the much smaller number, the results were highly concordant to the full simulation. Specifically, the entropy of the distribution was 3.93 bits and 7.25 bits at 10yr and 1yr bins, respectively. These results demonstrate that the generation time estimates are robust to the strata of the genealogical data. This is concordant with previous studies, which have shown that generation times show relatively little variation across human cultures and large time spans. A meta-analysis of recent (>1970) generation times across developed countries and a few hunter-gatherer societies showed a small difference of 1.5 years between the average generation time of well developed countries (30.1 years) to hunter-gatherer societies (28.6 years) (*29*). In addition, a recent study estimated the average generation time across the last 45,000 years using Neanderthal genomes and predicted an average of 28.1 years (*30*). Such small differences should not substantially affect our conclusions regarding the power of age as an identifier.

Direct analysis of a large set of relatives: Our data also allows to measure the year of birth differences in a large set of known distant relatives, as in the geographical analysis above. However, age analysis is more complicated when measured with recently born relatives due to ascertainment bias issues. First, the simulations above showed that for some types of relatives such as third cousins, the potential relative can be 90 years younger than the examined person, meaning that the relative is yet to be born, creating a censoring effect on our data. Second, in our previous studies with this dataset, we found that most individuals in our data came from the late 19th century. We were thus concerned that relatives ascertained from recently born individuals would disproportionally reflect these old cases and skew the age analysis.

As an alternative, we focused on historical data rather than recent data. We retrieved 1.2 million pairs of 2nd cousins and 1.7 million pairs of 3rd cousins that were born between 1800 to 1910 from the extensive family trees in the dataset of Geni.com, which was discussed in our previous publication (*19*). All of these pairs had exact birth date data and known birth locations. We found that the differences in the year of the actual 2nd and 3rd cousins were relatively similar to their simulations. For example, for 2nd cousins the entropy of the observed data was 6.04bits and 2.74bits vs. 6.17bits and 2.87bits in the simulations for 1yr resolution and 10yr resolution, respectively. For 3rd cousins the entropy of the observed data was 6.25bits and 2.95bits vs. 6.40bits and 3.09bits in the simulations for 1yr resolution and 10yr resolution, respectively. These small differences can stem from other types of ascertainment biases with the historical data or resemblance between relatives that induces reproduction at similar ages. Nevertheless, the overall consistency indicates that the simulation captures well the distribution of ages in these classes of relatives.

5. Identifying a 1000Genomes sample

We downloaded the joint VCF files of the 1000Genomes from the following link:
ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/. Next, we used the following shell script to extract variants for each genome:

```
#!/bin/bash
NA=$1

for chr in {1..22}
do
     bcftools query -f '%ID\t%CHROM\t%POS\t%REF\t[%TGT]\n' -s $NA -o $NA.genotypes.$chr
"ALL.chr"$chr".phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz" &

done
```

The script takes a sample name (e.g. NA12345) and extracts all variants. Next, we ran the following script to convert the variants into a DTC format:

```
#!/bin/sh
```

```
NA=$1
echo "Processing $NA"

echo "Merging all files..."
cat $NA.genotypes.* | awk '{print $2"\t"$3"\t"$3+1"\t"$0}' > $NA.all.genotypes.bed
echo "Finished merging all files"

echo "Retaining only DTC SNPs..."
bedtools intersect -a $NA.all.genotypes.bed -b DTC.bed > $NA.all.genotypes.bed.dtc
echo "Finished retaining only DTC SNPs"

echo "Sorting file..."
awk '{print $4"\t"$5"\t"$6"\t"$8}' $NA.all.genotypes.bed.dtc | uniq | sed -r 's/\|//'|
sort -k2 -n -k3 -n > $NA.all.genotypes.bed.snps.nearly_there
echo "File is sorted"

FINAL=$NA"_Genome.txt"
echo "Preparing final file ($FINAL)..."
grep '#' Example_Genome.txt > $FINAL
cat $NA.all.genotypes.bed.snps.nearly_there >> $FINAL
cat Example_Genome.txt | grep 'X' >> $FINAL
cat Example_Genome.txt | grep 'Y' >> $FINAL
cat Example_Genome.txt | grep 'MT' >> $FINAL
echo "Finished!"
```

DTC.bed is a file prepared from the raw data of a DTC provider that lists the chromosome number, position, and position+1 coordinates of each SNPs. Example_Genome.txt is a raw data file from a DTC provider. These two files can be prepared from online resources that list raw data from DTC providers, such as OpenSNP.org or the Personal Genomes Project. The scripts output a file formatted as a DTC-rendered genome.

Next, we processed a 1000Genomes sample that was identified in our previous study by inferring the surname of her husband (pedigree 3 in Fig. 3 of (*17*)). In our previous studies, our IRB determined that this process considered as Exempt Human Subject Research since the data is publicly available and we are not publicly reporting any identifiers of the sample. We uploaded the DTC rendered genome to GEDmatch and used the one-to-many comparison to search for matches. We then focused on two matches that had extensive pedigrees. We used the pedigrees to identify the ancestral couple that connects both of them and also the 1000Genomes sample. We then scanned descents of the ancestral couple and looked for a record that can match the 1000Genome project demographic identifiers. These included being a female in Utah that was alive in the year when the samples were collected, the year of birth, having parents that were alive in the year when the samples were collected, being married, and have exactly the same number of kids as the 1000Genomes sample. This process was time consuming due to the large number of descendants of the ancestral couple but we were eventually able to identify her via an obituary.

4. Cryptographic signatures

DTC genotyping files usually start with headers labeled with "#" sign that contains data about when the file was generated and what build of the genome it uses. Then, they follow with tabulated four columns that represent the SNP rsID, chromosome, position, and genotypes.

We suggest to add a cryptographic signature to the header file for compatibility with the current format and maximum usability. As an early prototype, we create a Python script that can sign and validate DTC files. Our script acts as a wrapper to a publicly available program called minisign that uses the EdDSA signature scheme. We note that the script should not be treated as a full implementation as it lacks key management. Rather, we envision that the script should be used to facilitate further discussions in the community regarding signing and validating files. Other options for cryptographic signature can include using GPG to sign, validate, and distribute the keys.

If adopted, cryptographic signatures will enable to block uploads of genome datasets of research participant to third party services. These files will lack the cryptographic signature of a valid DTC provider (or will have a cryptographic signature of a research genome center) and therefore will not be processed by third parties. As such, this method can block further attempts to identify 1000Genomes samples and other genomes currently stored in research projects.

We also posit that cryptographic signatures will substantially complicate the ability of an adversary to exploit long range familial searches to identify DNA samples. If only DTC providers can generate files processed by relative finders of third parties, the adversary will need to submit the sample to a valid DTC provider in order to obtain genotypes. A previous study reported that forensic samples can be genotyped with high throughput genotyping array using whole genome amplification (*32*). However, the call rates (SNPs that could be genotyped) were below 95% in nearly all (43 out of 44) types of samples that included hair, blood stains, and semen. These call rates are well below the typical call rates of DTC providers. For example, in scanning nearly 4000 23andMe and Ancestry samples in our database, the average no call rate was around 1% and always nearly always below 4%. The only exception were 4 samples with no call rates of >10% that are likely to be erroneous files of the users that uploaded the datasets to our system. Based on our experience, DTC providers usually consider no call rates of <95% as a failed test. Rather than reporting the partial results back to the customer, the provider typically asks for an another sample. This will greatly reduce the probability that the adversary can get a legitimately signed genotype file with a valid signature. Moreover, the <95% call rate reported in the literature is with laboratory procedures that were carefully calibrated for forensic samples. When sending the sample to a DTC provider, the adversary will have to dilute the sample DNA by the same lysis buffers used to collect and preserve saliva. For example, MyHeritage and FTDNA collect buccal swabs in tubes that contain lysis buffer of 500ul. Only a small fraction of the lysis buffer is subject to whole genome amplification. Thus, the sample will be further diluted before applied to the array, which is likely to further reduce the call rates. Additionally, beyond technological challenges, the adversary will have to face logistical and legal challenges. These include maintaining a shipping and receiving address in countries that are supported by DTC providers, have a valid credit card (that can reveal their identity), and agree to terms and services that prohibit sending samples without consent. Even if all of these challenges can be overcome, sending the samples to a DTC provider is still slower and more cumbersome option that working with a local lab. Thus, this procedure can slow the adversary and reduce the harm.

Supplemental Figures

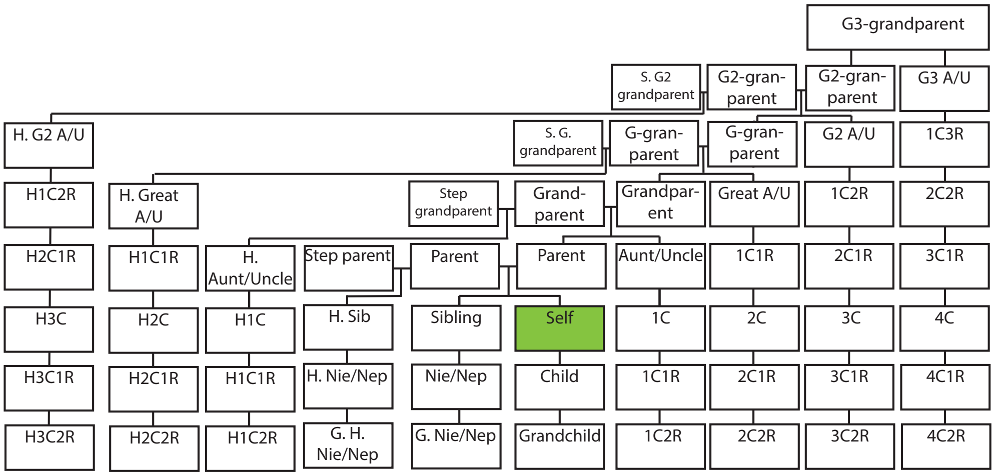


fig. S1: Examples of close and long-range genealogical relationships mentioned in this manuscript. Nie/Nep: Niece/Nephew; G2: Great-great; G3: Great-great-great; A/U: Aunt/Uncle. H: half. S: Step

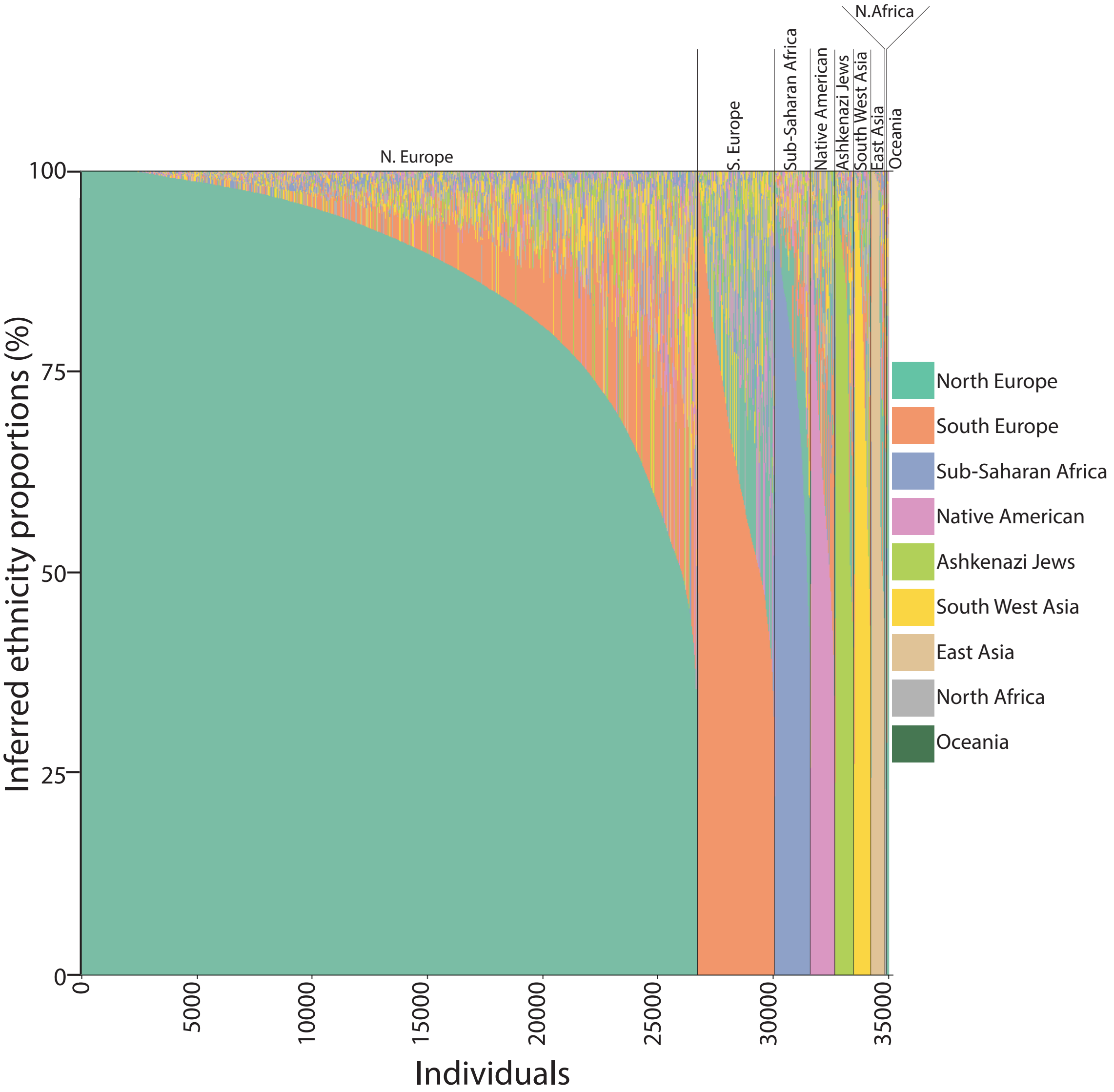


fig. S2: The genetic ethnicity of 36,000 users in the matching database. Each vertical line corresponds to a person and the Y-axis reflects the ethnicity composition from 0 to 100%. Colors denote the main ethnic groups in this analysis. The label for each group (top) is based on the major ethnicity in the group.

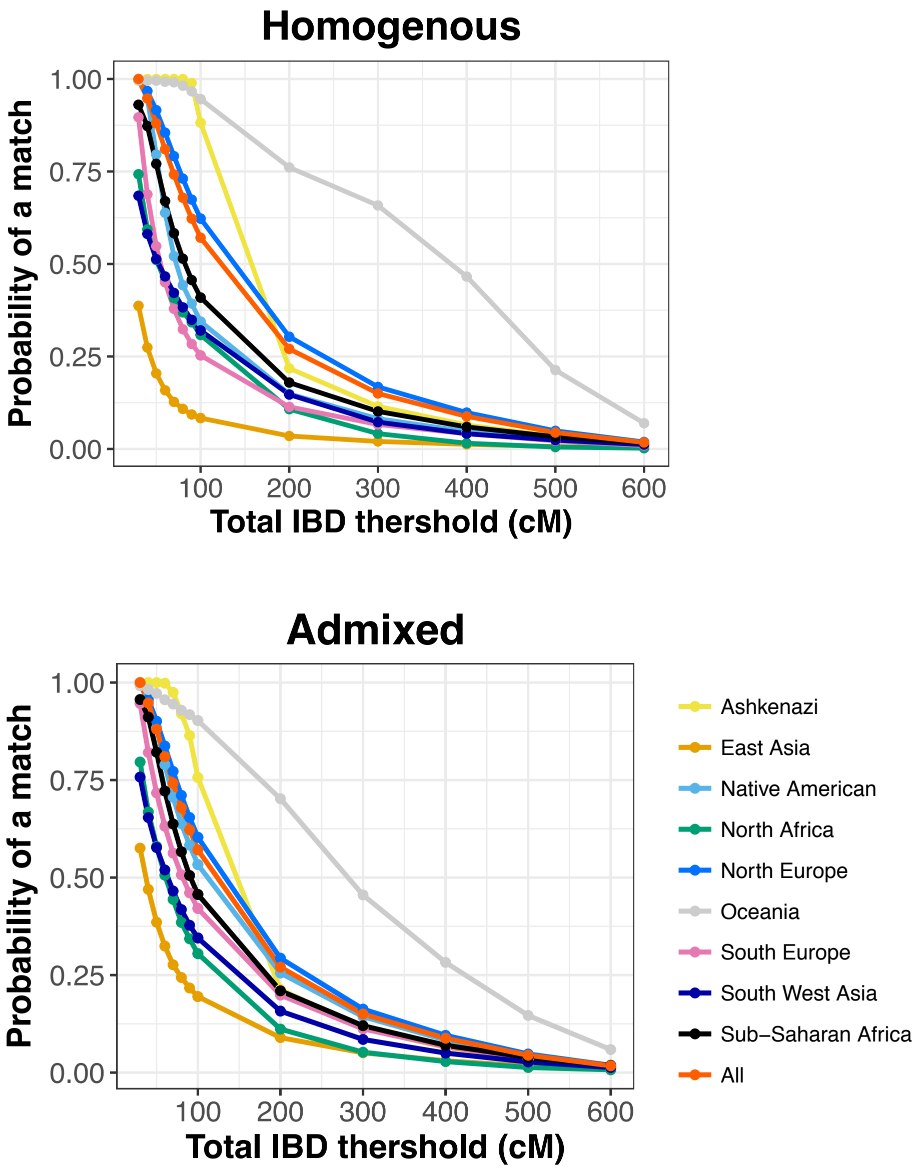


fig. S3: The probability of a match as a function of IBD threshold for various ethnicities. Each individual was compared to the entire database of 1.28M individuals as in Fig. 1A. We then stratified the results based on the primary genetic ethnicity of the individual. Homogenous: only individuals whose primary ethnicity consists is 80% of their ethnicity composition. Admixed: all individuals.

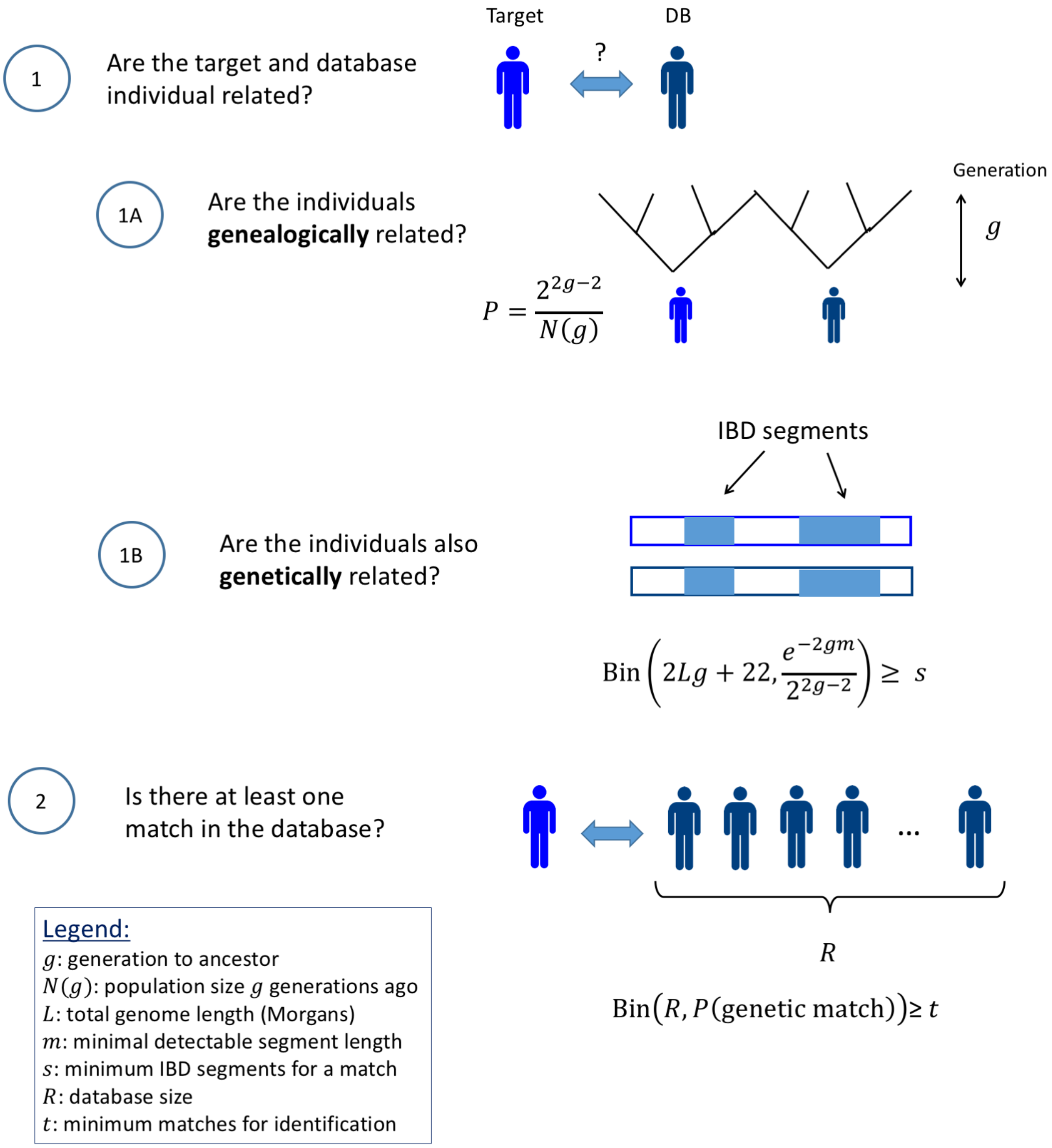

Target
DB
?
1
Are the target and database individual related?
1A
Are the individuals **genealogically** related?
Generation
g
$P = \frac{2^{2g-2}}{N(g)}$
IBD segments
1B
Are the individuals also **genetically** related?
$\mathrm{Bin}\left(2Lg + 22, \frac{e^{-2gm}}{2^{2g-2}}\right) \geq s$
2
Is there at least one match in the database?
...
R
$\mathrm{Bin}(R, P(\text{genetic match})) \geq t$
Legend:
g: generation to ancestor
$N(g)$: population size g generations ago
L: total genome length (Morgans)
m: minimal detectable segment length
s: minimum IBD segments for a match
R: database size
t: minimum matches for identification

fig. S4: Schematic illustration of our model for the probability to find a relative following a long range familial search. The model first evaluates the probability that the person of interest and the person in the database are genealogically related (1A). Then, it estimates the probability that these two individuals share enough IBD segments to be detected by the matching algorithm (1B). Finally, it calculates the probability of finding at least *t* matched individuals in a database size of R people.

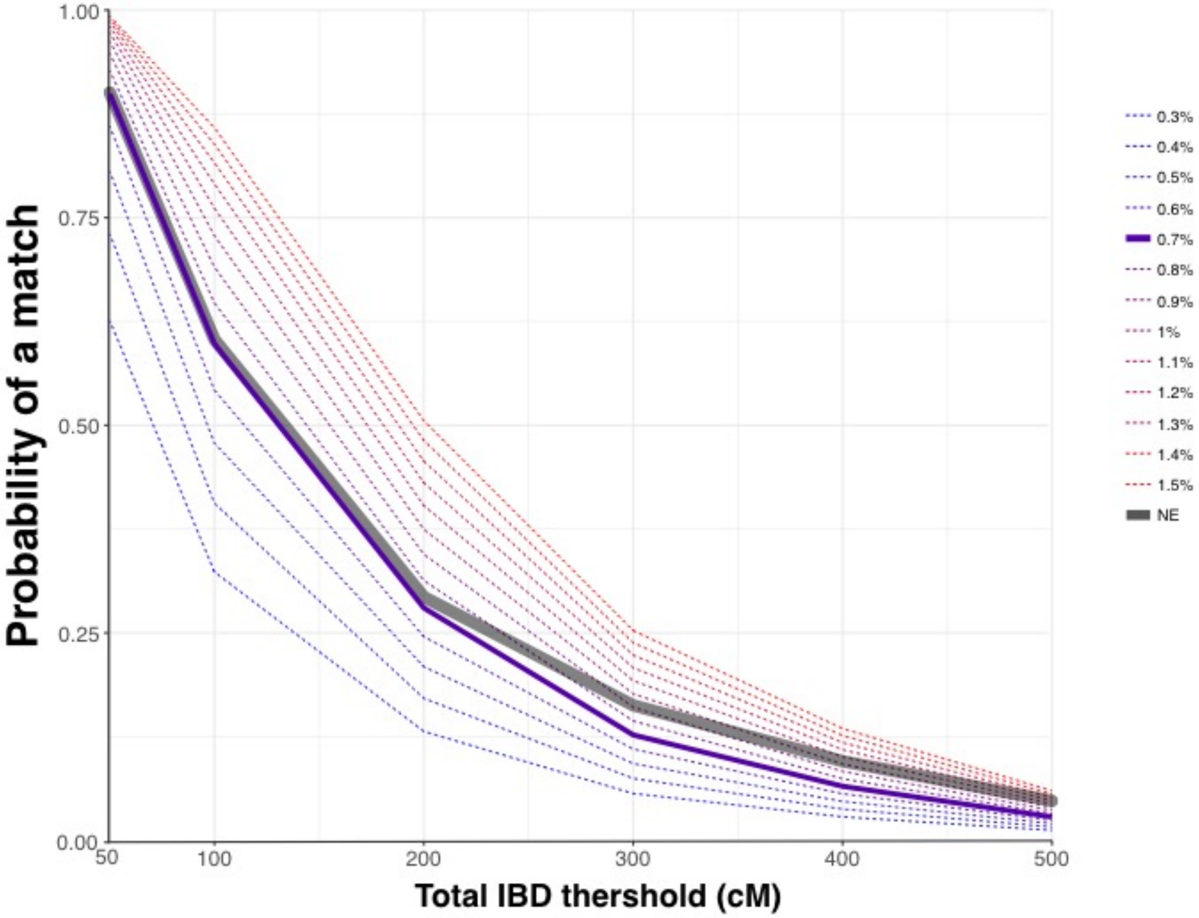
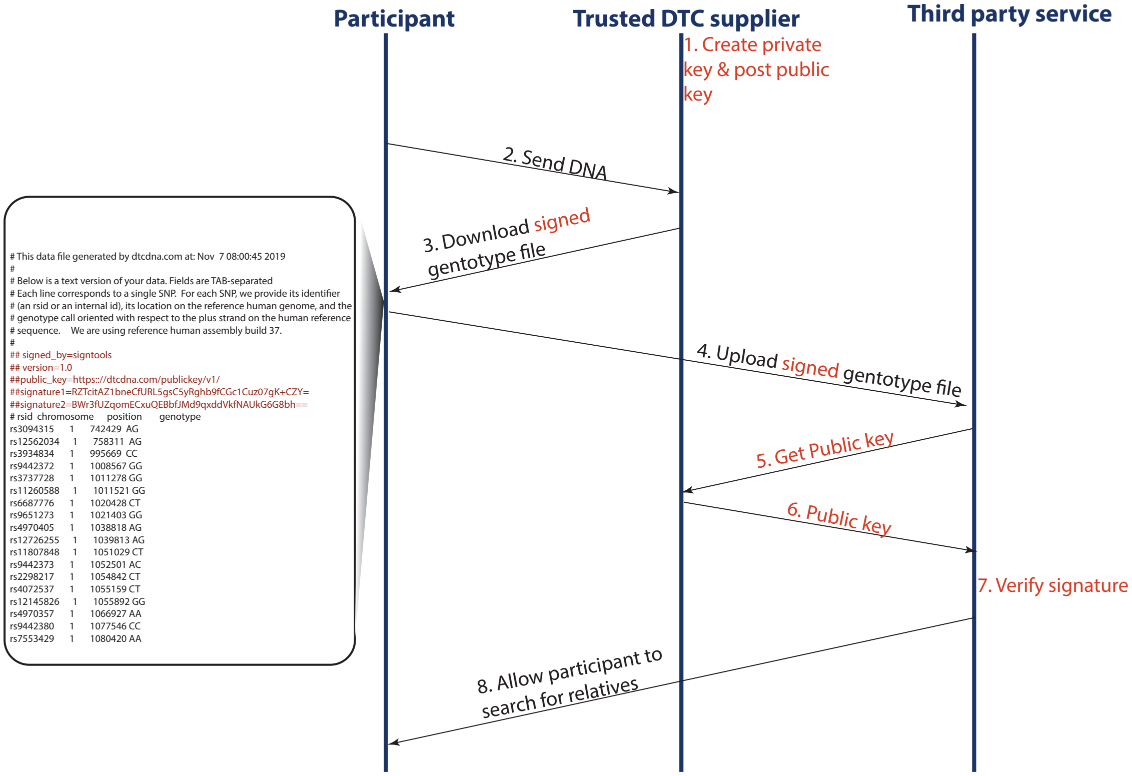


fig. S5: Theoretical model versus empirical results. The colored lines represent the theoretical model prediction for a probability of a match (i.e., finding at least one match in the database) as a function of the IBD threshold for various database sizes that cover between 0.3% to 1.5% of the population. The black line ("NE") depicts the empirical matching results of 967,418 individuals of primarily North European descent in our database. Based on US Census, there are 240M adult individuals in the US, 60% of which are primarily of European heritage. Thus, 967K individuals cover (967K/(0.6*240M) ≈) 0.7% of the population. The 0.7% line is highlighted versus the empirical results and shows very good consistency, especially for 250cM and below, which correspond to 2C relationships and above.


Participant
Trusted DTC supplier
Third party service
1. Create private key & post public key
2. Send DNA
3. Download signed gentotype file
4. Upload signed gentotype file
5. Get Public key
6. Public key
7. Verify signature
8. Allow participant to search for relatives

```
# This data file generated by dtcdna.com at: Nov  7 08:00:45 2019
#
# Below is a text version of your data. Fields are TAB-separated
# Each line corresponds to a single SNP.  For each SNP, we provide its identifier
# (an rsid or an internal id), its location on the reference human genome, and the
# genotype call oriented with respect to the plus strand on the human reference
# sequence.     We are using reference human assembly build 37.
#
## signed_by=signtools
## version=1.0
##public_key=https:://dtcdna.com/publickey/v1/
##signature1=RZTcitAZ1bneCfURL5gsC5yRghb9fCGc1Cuz07gK+CZY=
##signature2=BWr3fUZqomECxuQEBbfJMd9qxddVkfNAUkG6G8bh==
# rsid  chromosome     position       genotype
rs3094315      1      742429  AG
rs12562034      1      758311  AG
rs3934834      1      995669  CC
rs9442372      1      1008567 GG
rs3737728      1      1011278 GG
rs11260588      1      1011521 GG
rs6687776      1      1020428 CT
rs9651273      1      1021403 GG
rs4970405      1      1038818 AG
rs12726255      1      1039813 AG
rs11807848      1      1051029 CT
rs9442373      1      1052501 AC
rs2298217      1      1054842 CT
rs4072537      1      1055159 CT
rs12145826      1      1055892 GG
rs4970357      1      1066927 AA
rs9442380      1      1077546 CC
rs7553429      1      1080420 AA
```

fig. S6: The flow of data for validating files. The chart shows how third-party services can work together with trusted DTC suppliers in order to validate the authenticity of the data. Black: current flow of information. Red: added steps to authenticate the file. On the left is a snippet of a raw genotype file after signing.

Supplemental Tables

Service	Database size	DTC provider	Relative finder	Third-party support
23andMe	5M	●	●	
Ancestry	9M	●	●	
DNA.Land	100K		●	●
FTDNA	1M	●	●	●
GEDmatch	1M		●	●
LivingDNA	n/a	●		
MyHeritage	1.4M	●	●	●

Table S1: Vendors in consumer genomics, sorted lexicographically. Database sizes were taken from (*2*) on the basis of data available as of May 2018. DTC provider is a service that produces genomic information from biological material such as buccal swabs or saliva. Third-party support refers to services that allow upload of raw genomic data. The list includes only DTC providers or third-party services with relative finder mentioned in (*2*, *3*).

Case	Link	Reference
Buckskin Girl	https://www.forensicmag.com/news/2018/04/buck-skin-girl-case-break-success-new-dna-doe-project	(*32*)
Golden State Killer	https://www.theatlantic.com/science/archive/2018/04/golden-state-killer-east-area-rapist-dna-genealogy/559070/	(*33*)
Lyle Stevik	https://www.forensicmag.com/news/2018/05/dna-doe-project-ids-2001-motel-suicide-using-genealogy	(*34*)
William Earl Talbott II	https://www.washingtonpost.com/news/morning-mix/wp/2018/05/21/a-genealogy-website-used-to-crack-another-cold-case-police-say-this-one-a-1987-double-homicide/?utm_term=.02ce7e38237f	(*35*)
Joseph Newton Chandler III	https://www.washingtonpost.com/news/morning-mix/wp/2018/06/22/he-stole-the-identity-of-a-dead-8-year-old-police-now-want-to-know-what-he-was-hiding-from/?utm_term=.3f98ef680528	(*36*)
Gary Hartman	https://www.cnn.com/2018/06/22/us/cold-case-killing-1986/index.html	(*37*)
Raymond "DJ Freez" Rowe	https://lancasteronline.com/news/local/raymond-dj-freez-rowe-arrested-for-murder-of-schoolteacher-christy/article_f05a2ee4-78b2-11e8-ad10-4382ef42f96d.html	(*38*)
James Otto Earhart	https://www.kagstv.com/article/news/local/brazos-county-sheriff-announces-suspect-in-decades-old-murder-of-virginia-freeman/499-567341120	(*39*)
John D. Miller	https://www.washingtonpost.com/local/public-safety/in-decades-old-crimes-considered-all-but-unsolvable-genetic-genealogy-brings-flurry-of-arrests/2018/07/16/241f0e6a-68f6-11e8-bf8c-f9ed2e672adf_story.html?utm_term=.3613195e3f70	(*40*)
Matthew Dusseault/Tyler Grenon	http://www.providencejournal.com/news/20180727/how-dna-and-tattoo-led-to-charges-in-cold-ri-murder-case	(*41*)
Spencer Glen Monnett	https://www.thespectrum.com/story/news/2018/07/28/79-year-old-woman-raped-assaulted-her-st-george-home/855583002/	(*42*)
Darold Wayne Bowden	https://www.nytimes.com/2018/08/23/us/ramsey-street-rapist-dna.html	(*43*)
Michael F. Henslick	http://www.news-gazette.com/news/local/2018-08-29/cassano-case-suspect-lived-within-blocks-victim.html	(*44*)

Table S2: Links to announcement of long range familial searches for law enforcement cases.

Main DNA ethnicity	Percentage
North Europe	76.3%
South Europe	9.5%
Sub-Saharan Africa	4.5%
Native American	2.9%
Ashkenazi Jewish	2.4%
South/West Asia	2.1%
East Asia	1.7%
North Africa	0.3%
Oceania	0.2%

Table S3: The fraction of individuals in each major genetic ethnicity for the 1.28M individuals in our dataset.

Model prediction	1C2R	2C1R	3C	2C2R	2C
#cases	33974	32432	13522	58018	7712
>100km	0.679049	0.627467	0.606567	0.664897	0.563797
>100miles	0.5994	0.539097	0.543854	0.594746	0.489627
>200km	0.566021	0.508263	0.522556	0.570616	0.451504

Table S4: The probability that a relative is found outside of a 100km,100 miles, and 200km range from a match.

References and Notes

1. R. Khan, D. Mittelman, Consumer genomics will change your life, whether you get tested or not. *Genome Biol.* **19**, 120 (2018). doi:10.1186/s13059-018-1506-1 Medline

2. L. Larkin, Autosomal DNA testing comparison chart, The DNA Geek; http://thednageek.com/dna-tests/.

3. S. C. Nelson, S. M. Fullerton, "Bridge to the literature"? Third-party genetic interpretation tools and the views of tool developers. *J. Genet. Couns.* **27**, 770–781 (2018). Medline

4. A. Gusev, J. K. Lowe, M. Stoffel, M. J. Daly, D. Altshuler, J. L. Breslow, J. M. Friedman, I. Pe'er, Whole population, genome-wide mapping of hidden relatedness. *Genome Res.* **19**, 318–326 (2009). doi:10.1101/gr.081398.108 Medline

5. C. D. Huff, D. J. Witherspoon, T. S. Simonson, J. Xing, W. S. Watkins, Y. Zhang, T. M. Tuohy, D. W. Neklason, R. W. Burt, S. L. Guthery, S. R. Woodward, L. B. Jorde, Maximum-likelihood estimation of recent shared ancestry (ERSA). *Genome Res.* **21**, 768–774 (2011). doi:10.1101/gr.115972.110 Medline

6. B. M. Henn, L. Hon, J. M. Macpherson, N. Eriksson, S. Saxonov, I. Pe'er, J. L. Mountain, Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. *PLOS ONE* **7**, e34267 (2012). doi:10.1371/journal.pone.0034267 Medline

7. International Society of Genetic Genealogy Wiki, Success stories (2018); https://isogg.org/wiki/Success_stories.

8. Y. Erlich, A. Narayanan, Routes for breaching and protecting genetic privacy. *Nat. Rev. Genet.* **15**, 409–421 (2014). doi:10.1038/nrg3723 Medline

9. J. Ge, R. Chakraborty, A. Eisenberg, B. Budowle, Comparisons of familial DNA database searching strategies. *J. Forensic Sci.* **56**, 1448–1456 (2011). doi:10.1111/j.1556-4029.2011.01867.x Medline

10. N. A. Garrison, R. V. Rohlfs, S. M. Fullerton, Forensic familial searching: Scientific and social implications. *Nat. Rev. Genet.* **14**, 445 (2013). doi:10.1038/nrg3519 Medline

11. J. Kim, D. Mammo, M. B. Siegel, S. H. Katsanis, Policy implications for familial searching. *Investig. Genet.* **2**, 22 (2011). doi:10.1186/2041-2223-2-22 Medline

12. M. Gafni, "Here's the 'open-source' genealogy DNA website that helped crack the Golden State Killer case," *Mercury News*, 26 April 2018; www.mercurynews.com/2018/04/26/ancestry-23andme-deny-assisting-law-enforcement-in-east-area-rapist-case/.

13. J. Jouvenal, "To find alleged Golden State Killer, investigators first found his great-great-great-grandparents," *Washington Post*, 30 April 2018; www.washingtonpost.com/local/public-safety/to-find-alleged-golden-state-killer-investigators-first-found-his-great-great-great-grandparents/2018/04/30/3c865fe7-dfcc-4a0e-b6b2-0bec548d501f_story.html?utm_term=.6ff5cff1630e.

14. P. Aldhous, "DNA data from 100 crime scenes has been uploaded to a genealogy website—just like the Golden State Killer," *BuzzFeed*, 17 May 2018; www.buzzfeed.com/peteraldhous/parabon-genetic-genealogy-cold-cases?utm_term=.tkKXDVOWq#.yyz8oGQWd.

15. See supplementary materials.

16. B. T. Bettinger, The Shared cM Project – Version 3.0 (2017); https://thegeneticgenealogist.com/wp-content/uploads/2017/08/Shared_cM_Project_2017.pdf.

17. M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, Y. Erlich, Identifying personal genomes by surname inference. *Science* **339**, 321–324 (2013). doi:10.1126/science.1229566 Medline

18. J. Yuan, A. Gordon, D. Speyer, R. Aufrichtig, D. Zielinski, J. Pickrell, Y. Erlich, DNA.Land is a framework to collect genomes and phenomes in the era of abundant genetic information. *Nat. Genet.* **50**, 160–165 (2018). doi:10.1038/s41588-017-0021-8 Medline

19. J. Kaplanis, A. Gordon, T. Shor, O. Weissbrod, D. Geiger, M. Wahl, M. Gershovits, B. Markus, M. Sheikh, M. Gymrek, G. Bhatia, D. G. MacArthur, A. L. Price, Y. Erlich, Quantitative analysis of population-scale family trees with millions of relatives. *Science* **360**, 171–175 (2018). Medline

20. J. Warren, R. Reboussin, R. R. Hazelwood, A. Cummings, N. Gibbs, S. Trumbetta, Crime scene and distance correlates of serial rape. *J. Quant. Criminol.* **14**, 35–59 (1998). doi:10.1023/A:1023044408529

21. H. Han, C. Otto, A. K. Jain, "Age estimation from face images: Human vs. machine performance," paper presented at the 6th International Association for Pattern Recognition (IAPR) International Conference on Biometrics (ICB), Madrid, Spain, 4 to 7 June 2013.

22. Department of Health and Human Services, Federal policy for the protection of human subjects. *Fed. Regist.* **82**, 7149–7274 (2017).

23. N. Ram, C. J. Guerrini, A. L. McGuire, Genealogy databases and the future of criminal investigation. *Science* **360**, 1078–1079 (2018). doi:10.1126/science.aau1083 Medline

24. D. J. Bernstein, N. Duif, T. Lange, P. Schwabe, B.-Y. Yang, High-speed high-security signatures. *J. Cryptogr. Eng.* **2**, 77–89 (2012). doi:10.1007/s13389-012-0027-1

25. V. Shchur, R. Nielsen, On the number of siblings and *p*-th cousins in a large population sample. bioRxiv 145599 [Preprint]. 3 May 2018. https://doi.org/10.1101/145599.

26. J. Wakeley, L. King, B. S. Low, S. Ramachandran, Gene genealogies within a fixed pedigree, and the robustness of Kingman's coalescent. *Genetics* **190**, 1433–1445 (2012). doi:10.1534/genetics.111.135574 Medline

27. P. Ralph, G. Coop, The geography of recent genetic ancestry across Europe. *PLOS Biol.* **11**, e1001555 (2013). doi:10.1371/journal.pbio.1001555 Medline

28. D. Edge, G. Coop, "How lucky was the genetic investigation in the Golden State Killer case?" *gcbias*, 7 May 2018; https://gcbias.org/2018/05/07/how-lucky-was-the-genetic-investigation-in-the-golden-state-killer-case/.

29. J. N. Fenner, Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am. J. Phys. Anthropol.* **128**, 415–423 (2005). doi:10.1002/ajpa.20188 Medline

30. P. Moorjani, S. Sankararaman, Q. Fu, M. Przeworski, N. Patterson, D. Reich, A genetic method for dating ancient genomes provides a direct estimate of human generation interval in the

last 45,000 years. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 5652–5657 (2016). doi:10.1073/pnas.1514696113 Medline

31. G. C. Kennedy, “A SNP-based microarray technology for use in forensic applications,” Final Technical Report (National Criminal Justice Reference Service, 2008); www.ncjrs.gov/pdffiles1/nij/grants/223977.pdf.

32. S. Augenstein, “‘Buck Skin Girl’ case break is success of new DNA Doe project,” *Forensic*, 16 April 2018; www.forensicmag.com/news/2018/04/buck-skin-girl-case-break-success-new-dna-doe-project.

33. S. Zhang, “How a genealogy website led to the alleged Golden State Killer,” *The Atlantic*, 27 April 2018; www.theatlantic.com/science/archive/2018/04/golden-state-killer-east-area-rapist-dna-genealogy/559070/.

34. S. Augenstein, “DNA Doe Project IDs 2001 motel suicide, using genealogy,” *Forensic*, 9 May 2018; www.forensicmag.com/news/2018/05/dna-doe-project-ids-2001-motel-suicide-using-genealogy.

35. M. Flynn, “A genealogy website helps crack another cold case, police say, this one a 1987 double homicide,” *Washington Post*, 21 May 2018; www.washingtonpost.com/news/morning-mix/wp/2018/05/21/a-genealogy-website-used-to-crack-another-cold-case-police-say-this-one-a-1987-double-homicide/?utm_term=.b19f059ff197.

36. K. Swenson,” He stole the identity of a dead 8-year-old. Police want to know what he was hiding from,” *Washington Post*, 22 June 2018; www.washingtonpost.com/news/morning-mix/wp/2018/06/22/he-stole-the-identity-of-a-dead-8-year-old-police-now-want-to-know-what-he-was-hiding-from/?utm_term=.3f98ef680528.

37. R. Ellis, “DNA on napkin used to crack 32-year-old cold case, police say,” CNN, 24 June 2018; https://edition.cnn.com/2018/06/22/us/cold-case-killing-1986/index.html.

38. J. Hawkes, T. Knapp, “Raymond ‘DJ Freez’ Rowe arrested for 1992 killing of schoolteacher Christy Mirack,” *LNP Lancaster Online*, 25 June 2018; https://lancasteronline.com/news/local/raymond-dj-freez-rowe-arrested-for-murder-of-schoolteacher-christy/article_f05a2ee4-78b2-11e8-ad10-4382ef42f96d.html.

39. J. O’Brien, K. Bowen, V. Croix, “Brazos County Sheriff announces suspect in decades-old murder of Virginia Freeman,” KAGS, 25 June 2018; www.kagstv.com/article/news/local/brazos-county-sheriff-announces-suspect-in-decades-old-murder-of-virginia-freeman/499-567341120.

40. J. Jouvenal, “The unlikely crime-fighter cracking decades-old murders? A genealogist,” *Washington Post*, 16 July 2018; www.washingtonpost.com/local/public-safety/in-decades-old-crimes-considered-all-but-unsolvable-genetic-genealogy-brings-flurry-of-arrests/2018/07/16/241f0e6a-68f6-11e8-bf8c-f9ed2e672adf_story.html?utm_term=.cbb1640c0d4b.

41. A. Milkovits, “How DNA and a tattoo led to charges in cold R.I. murder case,” *Providence Journal*, 27 July 2018; www.providencejournal.com/news/20180727/how-dna-and-tattoo-led-to-charges-in-cold-ri-murder-case.

42. D. DeMille, “Arrest made in home invasion rape of elderly St. George woman,” *The Spectrum*, 28 July 2018; www.thespectrum.com/story/news/2018/07/28/79-year-old-woman-raped-assaulted-her-st-george-home/855583002/).

43. J. Fortin, “In serial rape case that stumped police, genealogy database leads to arrest,” *New York Times*, 23 August 2018; www.nytimes.com/2018/08/23/us/ramsey-street-rapist-dna.html.

44. M. Schenk, “Cassano case: Suspect lived within blocks of victim,” *The News-Gazette*, 29 August 2018; www.news-gazette.com/news/local/2018-08-29/cassano-case-suspect-lived-within-blocks-victim.html.