Received: 17 May 2022 | Revised: 16 August 2022 | Accepted: 16 August 2022

DOI: 10.1111/1556-4029.15127

JOURNAL OF FORENSIC SCIENCES AAFS

ORIGINAL PAPER

CRIMINALISTICS

Analysis of the genealogy process in forensic genetic genealogy

Mine Su Ertürk PhD[1] | Colleen Fitzpatrick PhD[2] | Margaret Press PhD[3] | Lawrence M. Wein PhD[1]

[1]Graduate School of Business, Stanford University, Stanford, California, USA

[2]Identifinders, Fountain Valley, California, USA

[3]DNA Doe Project, Sebastopol, California, USA

Correspondence
Lawrence M. Wein, Graduate School of Business, Stanford University, Stanford, CA 94305, USA.
Email: lwein@stanford.edu

Abstract

The genealogy process is typically the most time-consuming part of—and a limiting factor in the success of—forensic genetic genealogy, which is a new approach to solving violent crimes and identifying human remains. We formulate a stochastic dynamic program that—given the list of matches and their genetic distances to the unknown target—chooses the best decision at each point in time: which match to investigate (i.e., find its ancestors and look for most recent common ancestors between the match and the target), which set of potential most recent common ancestors to descend from (i.e., find its descendants, with the goal of identifying a marriage between the maternal and paternal sides of the target's family tree), or whether to terminate the investigation. The objective is to maximize the probability of finding the target minus a cost associated with the expected size of the final family tree. We estimate the parameters of our model using data from 17 cases (eight solved, nine unsolved) from the DNA Doe Project. We assess the Proposed Strategy using simulated versions of the 17 DNA Doe Project cases, and compare it to a Benchmark Strategy that ranks matches by their genetic distance to the target and only descends from known common ancestors between a pair of matches. The Proposed Strategy solves cases ≈10 − fold faster than the Benchmark Strategy, and does so by aggressively descending from a set of potential most recent common ancestors between the target and a match even when this set has a low probability of containing the correct most recent common ancestor. Our analysis provides a mathematical foundation for improving the genealogy process in forensic genetic genealogy.

KEYWORDS
family trees, forensic genetic genealogy, investigative strategy models, performance analysis, probabilistic analysis, stochastic dynamic programming

Highlights

- We model and analyze the genealogy process in forensic genetic genealogy as a stochastic dynamic program.

Presented at the International Symposium for Human Identification, September 13–16, 2021, in Lake Buena Vista, FL; the Scientific Working Group for DNA Analysis Methods, October 26, 2021; the American Society of Criminology Annual Meeting, November 17-20, 2021, in Chicago, IL; and the 74th Annual Scientific Conference of the American Academy of Forensic Sciences, February 21-26, 2022, in Seattle, WA.

- We quantify the difficulty of a forensic genetic genealogy case prior to performing the genealogy.
- We attempt to mathematically optimize the genealogy process.
- Cases can be solved faster by tracking the progress of finding the most recent common ancestors.

1 | INTRODUCTION

Since the April 2018 arrest in the Golden State Killer case [1], forensic genetic genealogy (abbreviated hereafter by FGG and also called investigative genetic genealogy) has emerged as an important tool to solve cold criminal cases and to identify unidentified human remains [2]. In this approach, a biological sample from a crime scene or from human remains is genotyped using a set of single-nucleotide polymorphisms (SNPs) and is typically performed at a private commercial laboratory. The SNP data are uploaded to third-party services (e.g., GEDmatch PRO, FamilyTreeDNA) that compare DNA data files from people who have tested with direct-to-consumer DNA testing companies on the same or similar SNP datasets. These third-party services generate a list of genetic relatives based on the amount of DNA shared with each match. Recent analysis [3, 4] has elucidated the genealogical–genetic tradeoff: as the genetic distance increases, the number of genetic relatives at this distance increases but the amount of shared genetic material—and hence the likelihood of genetic detection—decreases. These analyses predict the expected number of matches at each distance (e.g., third or fourth cousins) as a function of the database size of the third-party service.

However, nearly all of the FGG work occurs on the back end of the process [5]: how to identify the unknown individual (referred to here as the target) using information from the match list. A lucid description of the basic strategy is given in [6]: an ascending stage that finds common ancestors between pairs of matches in an attempt to find the unknown most recent common ancestors (MRCAs) between the unknown target and each investigated match (e.g., the MRCAs between two first cousins are their common grandparents), and a descending stage that looks for an intersection (e.g., a marriage between the mother's side and the father's side of the target's family tree) among the descendants of the common ancestors identified in the first stage. Although several insightful case studies have been documented (e.g., [6, 7]), there has been no detailed mathematical analysis of this genealogy process. More specifically, a typical target may generate a match list with several hundred third and fourth cousins, and it is not obvious how many matches, and which of these matches, to investigate, nor is it obvious how to optimally look for an intersection among their families.

In this study, we formulate the problem of finding an optimal genealogy strategy as a stochastic dynamic program, which is the standard approach to solving multi-period optimization problems under uncertainty [8]. The objective is to find the target as quickly as possible, by maximizing the probability of identifying the target minus a cost associated with the expected number of people in the final family tree, which we refer to as the expected workload; workload is used in lieu of time to find the target because we lack data that maps the former metric into the latter metric. We estimate the values of the model's parameters using data from 17 unidentified remains cases from the DNA Doe Project. We derive theoretical results about the structure of the optimal genealogy strategy, which guides the construction of the Proposed Strategy. We use computer simulation of the 17 cases to compare the performance of the Proposed Strategy to that of a Benchmark Strategy, which loosely represents how the genealogy process is typically performed.

2 | MATERIALS AND METHODS

2.1 | Model assumptions

2.1.1 | Simplifying assumptions

We begin with several simplifying assumptions: generations (denoted by g, where the target is in $g = 0$, the target's children are in $g = -1$ and the target's parents are in $g = 1$, etc.) are discrete and nonoverlapping, there is no endogamy (i.e., there are no marriages between people who are biologically related) and we ignore half and double relationships. We also do not use any information about the geographical location or ethnicity of any individuals.

2.1.2 | Probabilistic assumptions

The primitive probabilistic assumptions are given by three search probabilities denoted by p, q_a, and q_d, which are identification probabilities for a person in the match list, someone's parents, and someone's children, respectively. The probability that we can correctly identify (i.e., find their true name) someone on the match list is p; note that some people use an alias when using the third-party service. Because there was only one instance out of thousands in our data set where only one parent of an individual was identified, if we are searching for an individual's parents then we assume that either both or neither can be identified, which allows us to consider ancestral couples rather than individual ancestors. During the search process, given a node in the network (i.e., a leaf in the family tree) representing a known person or a couple, links emanating from this node will be investigated in an attempt to identify a person's parents or one or more of their children, which occurs with a specific probability (defined below by q_a for parents and q_d for children). Any

given link may be investigated multiple times (e.g., by different genealogists) throughout the search process. We assume that each link is sampled once with a given probability (q_a or q_d), and this random outcome holds for all subsequent investigations of this link. This assumption is consistent with a scenario where there is full coordination among all genealogists (e.g., a family tree is created jointly by a team of genealogists), genealogists vary in their skill level, and the most skilled genealogist is brought in to help with the investigation of a difficult link. That is, a child or parent is declared unidentified only after the most skilled team member has attempted and failed to identify them.

2.1.3 | Informational assumptions

In our model, the genealogists receive two types of information from the third-party service. The first type is a set of matches that can be investigated, each with a corresponding total centimorgan (cM) value, which quantifies the amount of shared DNA between the target and the match. Prior to the attempted identification of a match, we observe only its cM value and hence assume that the relationship of a match to the target is not directly observable. We construct a probability distribution of this relationship via a two-step process: a probabilistic mapping from cM value to an integer-valued distance (or degree of relatedness) between the target and the match [9], and a probabilistic mapping from distance to relationship [10] (§1.3 of Supporting Information); the probability distribution of the relationship given cM value can also be directly obtained via the Relationship Predictor tool [11]. We assume that if a match is successfully identified then the relationship r becomes observable; while this assumption does not strictly hold in practice, the birth year of an identified individual is typically known, which often provides an educated guess as to the relationship r. We note that the relationship information of a match is used only to determine the number of generations between the match and the still-to-be-determined MRCA couple of the match and the target.

The second type of information concerns the clustering of matches. GEDmatch's Autocluster tool [12] strives to group matches in the match list into clusters of matches that share the same common ancestors in a given generation. If the Autocluster tool is used at generation g, the matches are grouped into 2^{g-1} clusters. If $g = 2$, then the matches are grouped into a maternal cluster and a paternal cluster, whereas if $g = 3$, then the matches are grouped into four clusters representing the four grandparental lines. We refer the reader to [13–15] for further information on the Autocluster tool.

Within each generation $g = 1, 2, \ldots$, we index each of the ancestral couples of the target by $c \in \mathcal{C}_g = \{1, \ldots, 2^{g-1}\}$; for example, in generation $g = 2$, ancestral couple 1 is the target's paternal grandparents and ancestral couple 2 is the target's maternal grandparents. Our informational assumptions (§1.3 in Supporting Information) regarding ancestral couples (§1.1 in Supporting Information) and the grouping of matches are guided by GEDmatch's Autocluster tool [12]. In our model, we assume access to an idealized perfect Autocluster tool that can identify groups of matches if they descend from a common ancestor; because the matches are also related to the target, this common ancestral couple would also represent an ancestral couple of the target. Specifically, we assume that prior to an investigation of match i, the Autocluster tool allows us to observe which ancestral couple of the target each match descends from, although we cannot fully characterize the identity of the ancestral couple; for example, if matches 1 and 2 are first cousins to the target but unrelated to each other, we may know that match 1 descends from ancestral couple 1 of the target in generation 2 and match 2 from ancestral couple 2 of the target in generation 2, but we cannot determine which ancestral couple represents the paternal or maternal grandparents of the target without further investigation.

2.2 | Parameter estimation

Details of the parameter estimation procedure appear in §2 in Supporting Information. For each of the 17 cases in Table 1, the data (igg.xls in Supporting Information) include the set of matches and the corresponding cM values, whether or not each match was investigated, whether or not each investigated match was successfully identified, and whether or not the case was solved. Of the three identification probabilities (p, q_a, q_d), the probability p of identifying a match is straightforward given the total number of investigated matches and the total number of investigated matches that were successfully identified (§2.2 in Supporting Information, Table 1). Given that we have only initial cM data and final data on whether the case was solved, we were unable to jointly estimate q_a and q_d. As an upper bound, we set $q_d = 0.98$ (§2.3 in Supporting Information) based on a nonpaternity probability of 0.02 and assuming that not finding a known person's child is driven entirely by misattributed paternity. We also perform a sensitivity analysis with a more conservative estimate of $q_d = 0.90$, which is meant to represent a lower bound.

Given q_d, we use a simulation-based approach to estimate q_a. The approach simulates 500 random versions of each of the 17 cases. In each of these 8500 simulations, we randomly generate a family tree rooted in the target using a five-step procedure (§2.4 in Supporting Information). For each identified match in a case and its known cM value, we randomly assign a degree of relatedness to the target according to the derived probability distribution, and randomly assign the match to a node in the target's family tree that is consistent with this genetic distance. Then we randomly assign each edge in the family tree a binary ascending value with probability q_a and a binary descending value with probability q_d, and choose q_a so that 4000 of the 8500 simulated cases were solved under the Benchmark Strategy, which is consistent with eight of 17 DNA Doe Project cases being solved (§2.5 in Supporting Information). Our estimate of q_a is 0.60 when $q_d = 0.98$ and $q_a = 0.64$ when $q_d = 0.90$.

The only other parameters to estimate in our model are the mean number of children born to a couple in generation g, n_g; in

TABLE 1 Aggregate data for each case. W_{50} and W_{90} are the workload necessary to solve 50% and 90% of the 500 simulated realizations of each case, respectively. Cases are ordered according to their W_{50} values

Case number j	Case solved p^j	Number of available matches M	Largest total cM $\max_{i \in \{1,\ldots,M\}} s_i$	Number of investigated matches $\lvert\mathcal{M}\rvert$	Number of identified matches $\lvert\mathcal{M}_y\rvert$	W_{50} for benchmark strategy	W_{90} for benchmark strategy
1	Yes	633	240	34	29	12,795	29,615
2	Yes	313	280	107	44	15,673	42,897
3	Yes	2136	360	80	66	16,410	40,969
4	No	610	100	72	43	17,864	40,925
5	No	221	1550	31	29	18,039	80,754
6	No	545	80	246	25	18,240	52,740
7	No	5007	70	39	24	18,403	60,962
8	Yes	795	60	72	68	18,567	48,993
9	No	928	140	56	37	18,806	49,118
10	Yes	2000	120	56	20	18,955	51,078
11	No	1373	90	232	171	18,996	48,799
12	Yes	308	80	74	54	19,026	117,696
13	Yes	2417	170	199	149	19,071	51,504
14	No	5059	70	288	212	19,076	48,130
15	Yes	2134	120	31	10	19,370	46,507
16	No	509	80	86	22	19,475	62,636
17	No	4257	60	50	41	20,296	48,292
Total	8	29,245	–	1753	1044	–	–

our simulations, the actual number of children born to a couple in generation g is a Poisson random variable with parameter n_g. These parameters are derived using U.S. fertility rate data since 1800 [16], and assuming that a generation spans 25 years. We also assume that the target is 40 years old in 2020 (§2.1 in Supporting Information), although other ages can be easily accommodated.

2.3 | Stochastic dynamic program

2.3.1 | System state

The stochastic dynamic program (formulated in §1.2 in Supporting Information) is defined by the system state, the available actions, the objective function, and the transition probabilities (i.e., the probability distribution of the new state given a specific action in the current state) [8]. The system state at time t is of high dimension and contains three components. The first component contains a list of all matches that have been investigated through time $t-1$ (we use a discrete-time model), along with each match's cM value and the ancestral couple from which the match descends.

The main novelty of our model is how we represent the current status of each ancestral couple in the second component of the system state. In our model, the intermediate goal of investigating a match is to find the correct MRCA couple between the target and the match. In this investigative process of a match, the probability that we can identify someone on the match list is p, and given that a person or couple has been identified, the probability that we can identify both of their parents is q_a. At the end of this ascending investigation, some of the match's ancestral couples in the MRCA generation may not be identified, and some of the identified ancestral couples in the MRCA generation (referred to as potential MRCA couples) may not be the correct MRCA couple (e.g., the MRCA couple between the target and the match is the match's maternal grandparents, but the genealogists identify the match's paternal grandparents).

More specifically, the second component of the system state is given by the pair $(P, |L|) = \left\{\left(P_{g,c}, \left|L_{g,c}\right|\right)\right\}_{g\in\{1,\ldots,g_{\max}\}, c\in C_g}$, where $L_{g,c}$ is the list of potential MRCA couples associated with generation–ancestral couple pair (g, c), $P_{g,c}$ is the probability that one of these $|L_{g,c}|$ couples is the correct MRCA couple between the target and the match and coincides with the ancestral couple c of generation g (although there can be more than one match descending from the ancestral couple c, each of these matches has the same correct MRCA couple), and $g_{\max}$ denotes the highest generation that contains a MRCA couple. Note that we only keep track of the size $|L_{g,c}|$ of a list, not the list's individual entries.

Because we allow multiple descents from an ancestral couple throughout the genealogy process, the third component of the system state contains all generation–ancestral couple pairs for which at least one descending search has been performed through time $t-1$. For each of these generation–ancestral couple pairs, the state includes the list of all potential MRCA couples on which a descent

has been performed through time $t-1$ and the probability that this list contains the correct MRCA couple.

2.3.2 | Actions

The decision maker in the stochastic dynamic program can take one of three types of actions at each time t after observing the system state at time t: investigate a particular match among the matches that have yet to be investigated, descend from a list representing a particular ancestral couple in a particular generation, or stop the investigation (presumably because the future workload cost outweighs the likelihood of finding the target).

2.3.3 | Objective

The objective function in the stochastic dynamic program is to maximize the probability of finding the target minus the expected workload cost, which is the product of a cost (which can be thought of as a Lagrange multiplier for an expected workload constraint) and the expected total number of people in the final family tree.

2.3.4 | Transition probabilities

The most difficult aspects of the analysis (§1.3–1.4 in Supporting Information) are to compute the transition probabilities (i.e., given the current state at time t and a specific action taken at time t, the probability distribution of the new state at time $t+1$), and the running costs (i.e., the probability of solving the case at time t minus the additional expected workload cost incurred at time t) in §1.5 in Supporting Information.

2.3.5 | Structural results

The immense state space of the stochastic dynamic program appears to preclude a direct computation of the optimal strategy. However, in §1.6 in Supporting Information, we prove five propositions that partially characterize the structure of an optimal strategy; that is, we derive characteristics that the optimal strategy possesses. These structural properties are leveraged to construct the Proposed Strategy.

2.4 | Two strategies

2.4.1 | Proposed strategy

The Proposed Strategy is defined precisely in §1.7 in Supporting Information, where its links to the structural properties derived in §1.6 in Supporting Information are delineated. At each point in time, the Proposed Strategy's two-step algorithm decides whether to perform a descending action from a particular generation–ancestral couple pair, to perform an ascending action from a particular generation–ancestral couple pair (i.e., investigate a particular match), or to terminate the genealogy process without finding the target. The algorithm is based on the cost-effectiveness of each possible ascending and descending action, which is calculated using the transition probabilities and expected workload derived in the formulation of the stochastic dynamic program in §1.3–1.5 in Supporting Information. Because our problem formulation trades off the probability of finding the target and the expected workload, the cost-effectiveness in our setting takes the form of the probability of finding the target divided by the expected workload. The Proposed Strategy has two thresholds, θ_a and θ_d, whose values are determined via a search procedure (§2.6 in Supporting Information); these parameters are used in lieu of the Lagrange multiplier on the expected workload.

In the first step of the algorithm, we compute the descending cost-effectiveness for each generation–ancestral couple pair, which is the probability that the descending action finds the target (conditioned on the current state, which includes the state of each generation–ancestral couple pair just after its last descent) divided by the expected workload incurred by the descent (conditioned on the current state). If the largest cost-effectiveness value among all generation–ancestral couple pairs exceeds the threshold $\theta_{d,}$ then we descend from this maximal generation–ancestral couple pair. Otherwise, we go to the second step of the algorithm.

In the second step, within each generation–ancestral couple pair, uninvestigated matches are prioritized according to their total cM value. Because several ascending actions may be necessary to obtain a list of potential MRCA couples that can lead to a successful descending action, for each generation–ancestral couple pair, we first find the smallest number of uninvestigated matches who descend from this ancestral couple such that the expected state after performing the ascending actions on these matches results in a descending cost-effectiveness value that exceeds the threshold θ_d. Using this specified number of matches, we then compute the ascending cost-effectiveness for each generation–ancestral couple pair to be the probability of finding the target after ascending these matches and then descending from the generation–ancestral couple pair divided by the expected workload associated with these ascending and descending actions. If the largest ascending cost-effectiveness value among all generation–ancestral couple pairs exceeds the threshold $\theta_{a,}$ then we investigate the uninvestigated match with the highest cM value for this maximal generation–ancestral couple pair. Otherwise, the Proposed Strategy terminates without finding the target.

2.4.2 | Benchmark strategy

The Benchmark Strategy is meant to crudely represent current practice and has a single parameter n, which denotes the total number of ascending investigations that are performed. Given n,

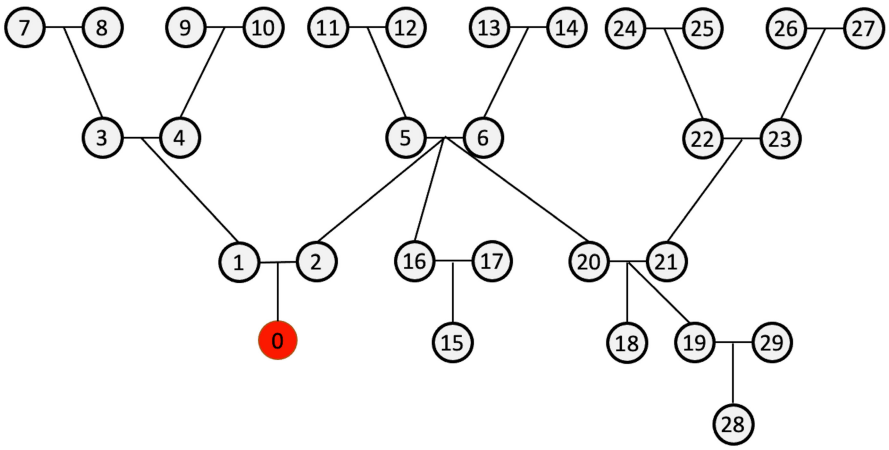

FIGURE 1 A sample family tree where nodes 15 and 18 are both first cousins of the target (node 0) and node 28 is a first cousin once removed

the Benchmark Strategy sequentially investigates matches that are ranked by the highest total cM, and performs a descending search whenever a common ancestral couple between a pair of matches is identified. The search is terminated after investigating n matches.

More specifically, an ascending investigation of a new match entails searching for the ancestors of this match in generations up to $g_{\max}$ and keeping track of those ancestral couples who have been successfully found. After each ascending action, the strategy verifies whether an ancestral couple identified during this ascending search has previously been identified during the ascending investigation of a different match. If indeed there is an intersection between the set of ancestral couples identified during the current ascending action and the ancestral couples identified previously, a descending search is initiated from this set of common ancestral couples. For example, suppose an ascending action is performed on node 18, who is a first cousin of the target, in the sample family tree in Figure 1, and suppose couples 5–6, 20–21, and 22–23 are identified. Suppose we then perform an ascending action on node 28, who is a first cousin once removed of the target, resulting in the identification of couples 5–6, 19–29, 20–21, and 22–23. At this point, couples 5–6, 20–21, and 22–23 have been identified twice and therefore a descending action is initiated from these three couples. In contrast, because couple 19–29 is identified only through node 28 (and not through node 18), this couple will not be included in this descending action.

Note that an ascending search under the Benchmark Strategy keeps track of common ancestral couples between pairs of matches rather than the MRCA couple between a match and the target. Because the MRCA of two matches does not necessarily coincide with the MRCA between a match and the target (see the Discussion for further details), the ascending search under the Benchmark Strategy investigates ancestors in generations up to $g_{\max}$ so that the common ancestors of the target and these two matches can actually be identified.

2.5 | Performance evaluation

The performance of the Benchmark Strategy and the Proposed Strategy are assessed by simulating each of the 17 DNA Doe Project cases 500 times, following a similar procedure used to estimate q_a (§2.4–2.5 in Supporting Information). On identical copies of the family tree generated by the simulation procedure, we run each strategy separately and truncate the simulation after 400 ascending actions in the Benchmark Strategy and 400 ascending plus descending actions in the Proposed Strategy (the Proposed Strategy was truncated in this manner in only 120 out of 8500 cases). In each simulation and at each time t, we record the action taken by the strategy and the following performance metrics: whether the target has been found, the cumulative ascending workload, and the cumulative descending workload. This information allows us to generate tradeoff curves of the proportion of the 8500 cases that are solved with a workload that is less than any specified value (Figure 2).

3 | RESULTS

3.1 | Parameter values

The parameter values appear in Table 2. These estimates suggest that it is easier to identify someone's children than someone's parents (not surprisingly, because parents may have lived in a time and place in which records were less accessible), and identifying a match is approximately as difficult as identifying someone's parents. In addition, Table 2 shows that lowering q_d in the sensitivity analysis is compensated for by an increase in q_a.

3.2 | Main results

The Proposed Strategy solves cases much more quickly than the Benchmark Strategy, which can be quantified by comparing the vertical or horizontal distance between the two curves in Figure 2. Comparing the vertical distance, we see that at a workload of 7500, the Proposed Strategy solves 94.3% of the cases compared to 4.4% for the Benchmark Strategy. Comparing the horizontal distance, we borrow from the definition of ID_{50} in the infectious dose literature and let W_x be the workload that solves x% of the 8500 simulated cases. For the Proposed Strategy, $W_{50} = 1646$ and $W_{90} = 5239$,

compared to W_{50} = 18, 226 and W_{90} = 51, 151 for the Benchmark Strategy; that is, relative to the Benchmark Strategy, the Proposed Strategy can solve 50% of the cases 11.1-fold faster and solve 90% of the cases 9.8-fold faster. It also achieves a higher probability of finding the target at the highest workloads that were evaluated: the Benchmark Strategy solves 93.3% of cases when the workload is 289,441, whereas the Proposed Strategy solves 97% of cases when the workload is 32,016 (Figure S3).

Although the aggregate curves in Figure 2 hide the variation across cases (Figure S4, where easier—that is, lower-numbered—cases have curves that are slightly to the upper left of the aggregate curve), all 34 curves in Figure S4 are increasing and concave. The relationship (Figure S5) between the number of investigated matches and the workload suggests that the curves in Figure 2 flatten out at approximately 185 investigated matches (corresponding to W = 7500) for the Proposed Strategy and 251 investigated matches (corresponding to W = 50k) for the Benchmark Strategy (these workload inflection points also vary by case, with easier cases having smaller inflection points).

3.3 | Behavior of the proposed strategy

The Proposed Strategy is quite robust with respect to the choice of threshold values: the results in Figure 2 use $(\theta_a, \theta_d) = (10^{-5}, 10^{-5})$, although the pairs $(10^{-4}, 10^{-4})$, $(10^{-4}, 10^{-5})$, $(10^{-5}, 10^{-4})$, $(10^{-6}, 10^{-6})$, and $(10^{-10}, 10^{-10})$ perform nearly identically (§2.6 in Supporting Information). Figure 3a reveals that the Proposed

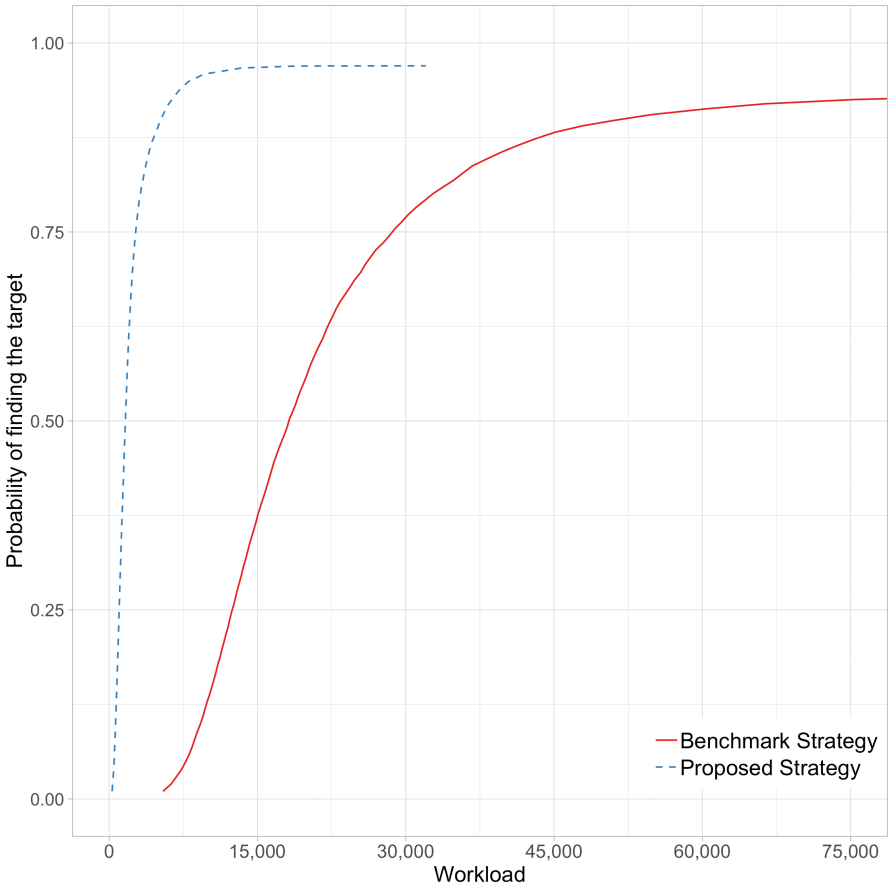

FIGURE 2 For the Proposed Strategy and the Benchmark Strategy, the proportion of 8500 simulated cases (500 for each of the 17 cases in Table 1 in the main text) that are solved with a workload less than the value on the horizontal axis

TABLE 2 Parameter values

Parameter	Description	Value
p	Probability that a match can be identified	0.596
q_a	Probability that parents can be identified	0.60 0.64 (S.A.)
q_d	Probability that a child can be identified	0.98 0.90 (S.A.)
n_g	Number of children per couple in generation g	2 if $g \leq 3$ $g - 1$ if $g > 3$

Abbreviation: S.A. stands for Sensitivity Analysis.

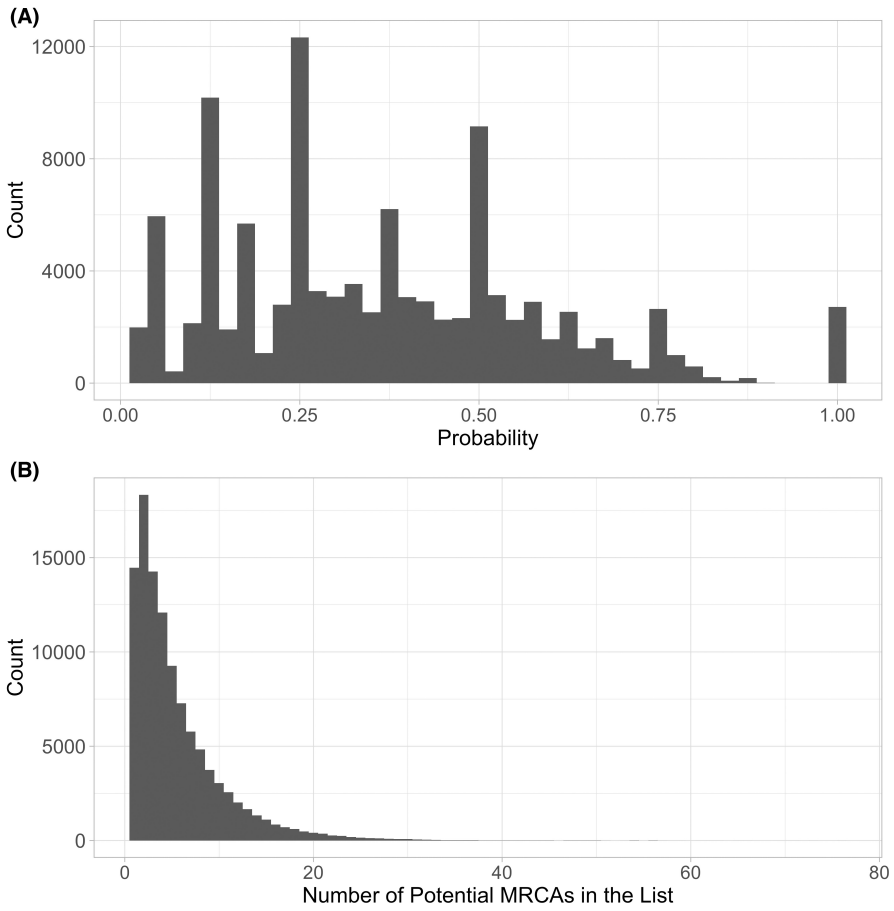

FIGURE 3 The histogram of (A) the probability that the list of potential MRCAs contains the correct MRCA, $P_{g,c}$, and (B) the number of potential MRCAs in the list, $|L_{g,c}|$, at the time of each descent from an ancestral couple over all 8500 simulations

Strategy solves cases so quickly by aggressively descending from ancestral couples when the probability that the list of potential MRCAs contains the correct ancestral couple is surprisingly small, with a mean of 0.36 over all descending actions in the 8500 simulated instances. The mean list size (i.e., the number of potential MRCAs in a list at the time of descent) is 5.5 (Figure 3b) and gets as large as 76.

Over all 8500 simulations, there are a total of 106,877 descents from 54,910 different ancestral couples, for an average of 12.6 descents per simulation and 1.9 descents per ancestral couple within a simulation. More generally, the sequence of actions taken by the Proposed Strategy varies across the 17 cases (Figure S6) and across the 500 iterations of a given case (Figure S7 displays the sequence of actions for the first 10 iterations of case 9, which is the case of median difficulty according to the W_{50} values), and appears to depend on the detailed network structure in a complicated way.

Because the Proposed Strategy is more aggressive at descending actions than the Benchmark Strategy, its average time of descent (i.e., time t if it is the tth action taken) is smaller (72 vs. 242) and its average workload per descending action (i.e., how many people are added to the family tree in a descending action) is smaller (176 vs. 2830). The great majority of the workload for both strategies is incurred during descents for all but the easiest cases under the Benchmark Strategy, and for the harder-than-average cases under the Proposed Strategy (Figure S8).

3.4 | Sensitivity analysis

The results (Figure S9) under the alternative set of parameters, (q_a, q_d) = (0.64, 0.90), are qualitatively similar to the results in Figure 2 (e.g., the W_{50} of the Benchmark Strategy is 9.8-fold larger than the W_{50} of the Proposed Strategy), with the main difference being that it is more difficult to solve cases under this alternative set of parameters (e.g., at high workloads, only 82% of cases are solved rather than 93–97%). The gap between the two strategies is somewhat smaller under this alternative set of parameters because the benefits stemming from the Proposed Strategy's aggressive descending actions are mitigated when q_d is lowered from 0.98 to 0.90.

3.5 | Performance evaluation of the benchmark strategy

Our simulation of the performance of the Benchmark Strategy offers a rapid way to assess the difficulty of a case prior to investigation. The 17 cases are ranked according to their W_{50} values (lowest is first) under the Benchmark Strategy in Table 1, which range from 12,795 to 20,296. We note that—perhaps not surprisingly given the problem complexity—there is no simple relationship between a case's W_{50} value and the third (the number of available matches) and fourth (the largest total cM value in the list) columns in Table 1: the linear regression $W_{50} = \beta_1 M + \beta_2 \max_{\{i=1,\ldots,M\}} s_i$ has an adjusted R^2 of

3.2×10^{-5} with neither coefficient being statistically significant. Note that the ranking of cases by W_{50} values differs from the ranking by W_{90} values. For example, case 12 has the largest W_{90} value, which is likely due to the small number of available matches coupled with the small cM value of the closest match.

To investigate the relationship between the W_{50} values and whether the case was actually solved by DNA Doe Project at the time the data were received, we note that the mean W_{50}-rank of the solved cases is 8.0 and the mean rank of the unsolved cases is 9.9. After we received the data in May 2020, cases 4 and 5 were subsequently solved by January 2022, which changes the mean W_{50}-ranks to 7.3 for solved cases and 11.4 for unsolved cases. On a related note, there is considerable intra-case variation in the workload required to (attempt to) solve the 500 simulated versions of each case; for example, for the case of median difficulty (case 9), the coefficient of variation (standard deviation divided by the mean) of the workload over the 500 simulated cases is 0.93 under the Proposed Strategy and 0.68 under the Benchmark Strategy (Figure S10).

4 | DISCUSSION

As the first attempt to provide a detailed mathematical model describing the genealogy process in FGG, perhaps the biggest contribution of this study is the general framework, which includes (a) framing the optimization problem as maximizing the probability of finding the target subject to a constraint on the expected workload, (b) tracking the investigative process using $(P_{g,c}, |L_{g,c}|)$ (i.e., the list of all potential MRCAs for each ancestral couple of the target, and the probability that the current list contains the correct MRCA between the target and a match), and (c) formulating and analyzing the problem as a stochastic dynamic program, which leads to the Proposed Strategy.

The Proposed Strategy performs much better than the Benchmark Strategy (Figure 2). This strong performance is achieved by tracking the progress of finding the correct MRCA couple between a match and the target via $(P_{g,c}, |L_{g,c}|)$ using the Autocluster tool and probabilistic information about the relationship between the target and the match, and then aggressively descending from ancestral couples where the probability that the list of potential MRCAs contains the correct MRCA is surprisingly small (averaging 0.36). In contrast, the Benchmark Strategy by construction descends only from common ancestral couples between two different matches without using the Autocluster tool or the relationship distribution.

To describe in more detail how these two strategies relate to one another, let M_1, M_2, and T denote two related matches and the target, and let $g(x, y)$ be the generation of the MRCA couple between individuals x and y. Then there are three cases that arise in the Benchmark Strategy, depending on the relative values of $g(M_1, M_2)$ and $\min\{g(M_1, T), g(M_2, T)\}$; that is, depending on the relationship between the two matches in comparison to their respective relationships with the target. When these two quantities are equal, then the MRCA couple between any pair of the three individuals—the two matches and the target—are the same ancestral couple, and the descending action in the Benchmark Strategy from the MRCA couple between the two matches is equivalent to descending from a generation–ancestral couple pair with state $(P, |L|) = (1, 1)$ in the Proposed Strategy. When $g(M_1, M_2) < \min\{g(M_1, T), g(M_2, T)\}$ (e.g., matches 1 and 2 are first cousins to each other, and both are second cousins to the target), then the MRCA couple between the two matches cannot be an ancestor of the target. In addition, without additional information (recall that the Benchmark Strategy does not make use of the Autocluster tool), we cannot determine whether any common ancestor of the two matches is also a common ancestor with the target. In this case, the descending action will involve multiple ancestral couples, some of which are unrelated to the target. Finally, when $g(M_1, M_2) > \min\{g(M_1, T), g(M_2, T)\}$, the Benchmark Strategy finds a common ancestral couple between the target and (at least) one of the matches, but it is not the most recent one, which can lead to inefficiencies (i.e., descending from ancestral couples in generation $g(M_1, M_2)$ that are unrelated to the target).

We highlight that the inefficiencies observed in the second and third cases cannot be avoided even with the use of the Autocluster tool, which would allow $g(M_1, M_2)$ to be observable. This is because any strategy that searches for the common ancestors between two matches (instead of the MRCA between a match and the target) has to ascend to higher generations in order to maintain a high success rate. Indeed, if the strategy were to ascend only up to $g(M_1, M_2)$ in order to reduce the workload, then (in the second case) the correct common ancestors between the matches and the target would not be included in the descending action; as a result, the success rate would decrease. Similarly, any such strategy cannot exactly determine which common ancestral couple between a pair of matches is a correct ancestral couple of the target, which would significantly reduce the workload incurred during descent. In contrast, by focusing on the MRCA between each match and the target, the Proposed Strategy is able to reduce the accumulated workload without compromising the success rate.

To tease out how the three key differences between the Proposed and Benchmark Strategies—the former (1) ascends to exactly the generation that contains the correct MRCA couple between the target and the match (via a combination of the relationship information and the Autocluster tool); (2) uses the Autocluster tool to prioritize among ascending actions based on their cost-effectiveness values, and (3) aggressively descends from ancestral couples based on their cost-effectiveness values—contribute to the performance gap, we consider an intermediate strategy, the Benchmark-g* Strategy, which descends whenever a generation–ancestral couple pair reaches the state $(P_{g,c}, |L_{g,c}|) = (1, 1)$ (hence, it explicitly considers the target, rather than focusing on the common ancestral couples between two matches), and otherwise continues ascending on the uninvestigated match with the highest total cM value. The Benchmark-g^* Strategy overcomes the first of the three differences noted above: it improves on the Benchmark Strategy by inferring how many generations to ascend to get to the generation of the correct MRCA between the match and the target, and the

Proposed Strategy improves on the Benchmark-g^* Strategy by descending more aggressively (i.e., does not wait until $(P_{g,c}, |L_{g,c}|) = (1, 1)$) and using the Autocluster tool to prioritize among ascending actions based on their cost-effectiveness values. Compared to the Benchmark Strategy, the Benchmark-g* Strategy is more efficient (e.g., the W_{50} of the Benchmark Strategy is 3.8-fold larger than the W_{50} of the Benchmark-g^* Strategy in Figure S11). This efficiency in the workload is driven by the fact that the Benchmark-g^* Strategy only ascends up to the generation containing the correct MRCA couple between each match and the target and hence avoids overshooting this generation, which would incur additional workload, particularly during subsequent descending actions. However, the Benchmark-g^* Strategy—in addition to being less efficient than the Proposed Strategy—asymptotes at a lower success rate because it descends only from the MRCA couples that have been identified with probability 1. These observations suggest that using a combination of the relationship information between the target and a match and the Autocluster tool can provide significant efficiency gains. Nevertheless, it is difficult to further tease out how the remaining gap between the Proposed and Benchmark-g^* Strategies is allocated between the two remaining differences. We highlight that simply adjusting the Proposed Strategy to descend when a generation–ancestral couple pair reaches the state (1, 1) would not suffice to tease out the remaining two differences because the ascending decisions also take into account whether a future descending action would be possible from the given generation–ancestral couple pair.

At least for the 17 DNA Doe Project cases, the Proposed Strategy is remarkably robust with respect to its threshold values: it essentially reduces to a single-threshold strategy, where excellent performance is generated by $\theta_a = \theta_d \in [10^{-10}, 10^{-4}]$ (§2.6 in Supporting Information). This robustness can be explained by the structure of the Proposed Strategy's decision rules in Steps 2 and 3 in §1.7 in Supporting Information. In particular, because both the ascending and descending actions are performed on the generation–ancestral couple pair or the match with the highest cost-effectiveness value, the value of the threshold only impacts which type of action (i.e., ascending, descending or ending) will be taken at a given time. For example, if the descending threshold is smaller than the highest descending cost-effectiveness value (i.e., $\theta_d < f_d[g^*, c^*]$), then a descending action will be taken from this generation–ancestral couple pair with the highest value. Furthermore, as the value of the threshold is increased, the strategy will continue to take the same action on the same generation–ancestral couple pair until a switching point where the condition $\theta_d < f_d(g^*, c^*)$ no longer holds. A similar reasoning applies to the ascending threshold as well. As a result, the evolution of the sample path under a given pair of threshold values affects the performance only by changing the type of action rather than changing the generation–ancestral couple pair or the specific match the action is applied to.

While we view the construction of the Proposed Strategy as the most interesting aspect of the analysis, our performance evaluation tool that simulates the performance of the Benchmark Strategy allows for the rapid comparison across cases (e.g., if a law enforcement agency has the budget to investigate only a subset of its cold cases, the W_{50} of each cold case can be quickly calculated, which can aid in deciding which cases to investigate) or within a case (e.g., the probability of finding the target vs. workload curve can be generated, which can help inform the decision of how much money or time to invest in any particular case) after the output from the third-party service (e.g., GEDmatch) is obtained.

4.1 | Limitations

Although our mathematical model of the genealogy process in FGG is somewhat complicated and captures most of the salient features of forensic genetic genealogy, there are several key characteristics that are omitted from the model. As noted earlier, we disallow half-relationships and endogamy, although they are present to varying degrees in many families. In addition, genealogists often use geographical location and ethnicity to help guide which ancestors and descendants to investigate (and which to discard). While the comparison of autosomal SNPs is the standard approach in FGG, short tandem repeats on the Y chromosome (Y-STRs) and SNPs on the Y chromosome can identify males of the same lineage, which can be used to prune the family tree during the ascending stage (e.g., [17–19]). We assume that the Autocluster tool is perfectly accurate, although this clustering information may degrade in practice after several generations; this assumption overstates the difference in performance between the two strategies. Also, the What Are The Odds (WATO) tool [20], which gives the conditional probability for possible tree locations of the target, can be valuable at certain stages of the genealogy process. Finally, our stochastic dynamic program assumes that if you descend from a generation–ancestral couple pair, then you descend from each of its $|L_{g,c}|$ potential MRCA couples. A more refined approach would allow the decision maker to descend from a particular potential MRCA couple in the list, but this would require a state space that is more detailed than $(P, |L|) = \left\{\left(P_{g,c}, \left|L_{g,c}\right|\right)\right\}_{g\in\{1,\ldots,g_{max}\}, c\in C_g}$. It may be worthwhile to try to incorporate these features into our framework.

While our estimated values for the identification probabilities p, q_a, and q_d of matches, parents and children should be useful as initial estimates, further refinement would be helpful. First, the 17 cases in Table 1 were not chosen randomly. Rather, the 17 cases are the result of a request that went out to all volunteer team leaders at the DNA Doe Project to provide us data on individual cases. While we are not in a position to assess whether there are systematic biases in these 17 cases, we do note that the DNA Doe Project has solved approximately half of their cases, which is not inconsistent with the fact that eight of these 17 cases had been solved at the time we received the data in May 2020, and 10 of the 17 cases have been solved by January 2022.

There are other limitations related to our parameter estimation. First and foremost, we could not directly estimate q_d from the DNA Doe Project data set. While $q_d \in [0.90, 0.98]$ seems like a reasonable range of exploration, future work should attempt a more rigorous estimate of this parameter, particularly because the asymptotic fraction of cases that are solved is somewhat sensitive to q_d (Figures S3 and S9). A natural model extension is to allow p, $q_{a,}$ and q_d to depend on the generation g (or, given the age of the target, to depend on the calendar year). Even though tracking down information about descendants is often more difficult if they never lived in the United States, these functions are not likely to be monotonically decreasing in g because the release of U.S. Census data is governed by the 72-year rule [21], which means that individual-level documentation is currently not available for any U.S. Census after 1940. Coupled with the widespread introduction of electronic records in the 1980s, the most difficult years in the twentieth century to obtain records is currently in the 1940–1985 range.

More generally, we had only initial (matches identified) and final (case solved or not) data to estimate q_a and q_d. However, more detailed data that included every identified and unidentified ancestor of every identified match, and every identified and unidentified descendant of every potential MRCA couple that was descended from, would allow for a straightforward generation-dependent estimate of q_a and q_d, respectively.

Because we do not have complete family trees from the DNA Doe Project, we are not in a position to compare the calculated workload in our model to the actual workload in the 17 cases. It is not clear whether our model overestimates or underestimates the actual workload incurred in these cases, and individual genealogists have different approaches and skill levels. It appears that genealogists sometimes searched farther in the past than need be, and focusing on common ancestors between two different matches can lead to more searching of descendants than necessary. On the other hand, their use of information about geographical location and ethnicity would certainly lead to some improvements in efficiency.

Moreover, we do not have data that allows us to map from workload (size of the family tree) to actual work time incurred by genealogists. Current FGG practice typically involves a law enforcement organization paying a specified amount of money for a specified amount of work. If the target is not identified after the specified amount of work, then the investigation is halted unless the law enforcement organization pays additional money, which they sometimes do not do. Without being able to predict actual workload (size of family tree) or link this quantity to the amount of time worked (or, alternatively, to know the proportion of cases that are solved during the initial specified amount of work), we are not in a position to estimate where the initial amount of time paid by law enforcement resides on the horizontal axis in Figure 2, and hence how many additional cases could be solved by adopting the Proposed Strategy.

Due to all the limitations discussed above, our results should be interpreted on a relative basis rather than an absolute basis. That is, the most robust findings are that the Proposed Strategy performs much better than the Benchmark Strategy, and most cases seem solvable but may require a very high workload (at least under the Benchmark Strategy) due to the decreasing returns of the investigative process (i.e., the concavity of the curves in Figure 2). Ultimately, genealogy is as much an art as a science, and the Proposed Strategy is intended to aid—and not to replace—the genealogists as they perform their important investigative work.

ACKNOWLEDGMENTS

The authors thank DNA Doe Project team leaders Robin Espensen, Ruth Foreman, Stacey Mitchell, Jenny Lecus, Anthony Lukas Redgrave, and Gina Wrather for sharing data from their cases. They also thank Vinita Cheepurupalli for processing the raw data.

FUNDING INFORMATION

Co-author Mine Su Ertürk PhD was funded by the Graduate School of Business, Stanford University.

CONFLICT OF INTEREST

Co-authors Colleen Fitzpatrick PhD and Margaret Press PhD are founders of Forensic Genetic Genealogy companies, and co-authors Mine Su Ertürk PhD and Lawrence M. Wein PhD have filed a provisional patent on the Proposed Strategy in this article.

ORCID

Mine Su Ertürk https://orcid.org/0000-0001-8096-0818
Lawrence M. Wein https://orcid.org/0000-0001-6125-0220

REFERENCES

1. Fuller T, Hauser C. Search for 'Golden state Killer' leads to arrest of ex-cop. New York, NY: New York Times; 2018. Accessed January 25, 2022. https://www.nytimes.com/2018/04/25/us/golden-state-killer-serial.html
2. Kling D, Phillips C, Kennett D, Tillmar A. Investigative genetic genealogy: current methods, knowledge and practice. Forensic Sci Int Genet. 2021;52:102474. https://doi.org/10.1016/j.fsigen.2021.102474
3. Erlich Y, Shor T, Pe'er I, Carmi S. Identity inference of genomic data using long-range familial searches. Science. 2018;362(6415):690–4. https://doi.org/10.1126/science.aau4832
4. Coop G. How lucky was the genetic investigation in the Golden state killer case? BioRxiv. 2019;531384. https://doi.org/10.1101/531384
5. Dowdeswell TL. Forensic genetic genealogy: a profile of cases solved. Forensic Sci Int Genet. 2022;58:102679. https://doi.org/10.1016/j.fsigen.2022.102679
6. Greytak EM, Moore C, Armentrout SL. Genetic genealogy for cold case and active investigations. Forensic Sci Int. 2019;299:103–13. https://doi.org/10.1016/j.forsciint.2019.03.039
7. Tillmar A, Fagerholm SA, Staaf J, Sjölund P, Ansell R. Getting the conclusive lead with investigative genetic genealogy–a successful case study of a 16 year old double murder in Sweden. Forensic Sci Int Genet. 2021;53:102525. https://doi.org/10.1016/j.fsigen.2021.102525

8. Bertsekas D. Dynamic programming and optimal control. Vol I. 4th ed. Nashua, NH: Athena Scientific; 2012.
9. Larkin L. The limits of predicting relationships using DNA. Accessed January 25, 2022. https://thednageek.com/the-limits-of-predicting-relationships-using-dna/
10. Bettinger B. The shared cM project: a demonstration of the power of citizen science. J Genet Geneal. 2016;8(1):38–42.
11. Nicholson B. Orogen: With population weights. Accessed March 1, 2022. https://dna-sci.com/tools/orogen-wtd/
12. Coakley L. Tips for using GEDmatch. Accessed January 25, 2022. https://genie1.com.au/tips-for-using-gedmatch/
13. Genetic Affairs. AutoCluster. Accessed January 25, 2022. https://www.geneticaffairs.com/features-autocluster.html
14. Leeds D. The Leeds Method. Accessed January 25, 2022. https://www.danaleeds.com/the-leeds-method/
15. MyHeritage Blog. Introducing AutoClusters for DNA matches. Accessed March 2, 2022. https://blog.myheritage.com/2019/02/introducing-autoclusters-for-dna-matches/
16. OurWorldInData.org. Fertility rate over the long-term, 1800–2017. Accessed January 25, 2022. https://ourworldindata.org/grapher/fertility-rate-complete-gapminder
17. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. Science. 2013;339(6117):321–4. https://doi.org/10.1126/science.1229566
18. Claerhout S, Roelens J, Van der Haegen M, Verstraete P, Larmuseau MH, Decorte R. Ysurnames? The patrilineal Y-chromosome and surname correlation for DNA kinship research. Forensic Sci Int Genet. 2020;44:102204. https://doi.org/10.1016/j.fsigen.2019.102204
19. Ge J, Budowle B. Forensic investigation approaches of searching relatives in DNA databases. J Forensic Sci. 2021;66(2):430–43. https://doi.org/10.1111/1556-4029.14615
20. dnapainter.com. What are the odds?. Accessed January 25, 2022. https://dnapainter.com/tools/probability
21. Heimlich R. The '72-year rule' governs release of census records. Washington, DC: Pew Research Center 2012. Accessed January 25, 2022. https://www.pewresearch.org/fact-tank/2012/04/09/the-72-year-rule-governs-release-of-census-records/

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Ertürk MS, Fitzpatrick C, Press M, Wein LM. Analysis of the genealogy process in forensic genetic genealogy. J Forensic Sci. 2022;67:2218–2229. https://doi.org/10.1111/1556-4029.15127