Cite as: Y. Erlich *et al.*, *Science*
10.1126/science.aau4832 (2018).

Identity inference of genomic data using long-range familial searches

Yaniv Erlich^{1,2,3,4*}, Tal Shor¹, Itsik Pe'er^{2,3}, Shai Carmi⁵

¹MyHeritage, Or Yehuda 6037606, Israel. ²Department of Computer Science, Fu Foundation School of Engineering, Columbia University, New York, NY, USA. ³Center for Computational Biology and Bioinformatics (C2B2), Department of Systems Biology, Columbia University, New York, NY, USA. ⁴New York Genome Center, New York, NY, USA. ⁵Braun School of Public Health and Community Medicine, The Hebrew University of Jerusalem, Jerusalem, Israel.

*Corresponding author. Email: erlichya@gmail.com

Consumer genomics databases have reached the scale of millions of individuals. Recently, law enforcement authorities have exploited some of these databases to identify suspects via distant familial relatives. Using genomic data of 1.28 million individuals tested with consumer genomics, we investigated the power of this technique. We project that about 60% of the searches for individuals of European-descent will result in a third cousin or closer match, which can allow their identification using demographic identifiers. Moreover, the technique could implicate nearly any US-individual of European-descent in the near future. We demonstrate that the technique can also identify research participants of a public sequencing project. Based on these results, we propose a potential mitigation strategy and policy implications to human subject research.

Consumer genomics has gained popularity (1). As of April 2018, more than 15 million people have undergone direct-to-consumer (DTC) autosomal genetic tests, with about 7 million kits sold in 2017 alone (2). Nearly all major DTC providers use dense genotyping arrays that probe around ~700,000 genomics variants and let participants to download their raw genotype files in a plain text format. This has led to the advent of third-party services, such as DNA.Land and GEDmatch, which allow participants to upload their raw genotype files for further analysis (table S1) (3). Nearly all of these services offer to find genetic relatives by locating identity-by-descent (IBD) segments that can indicate a shared ancestor. Finding genetic relatives can accurately link even distant relatives, such as 2nd or 3rd cousins (4–6) (fig. S1) and has led to multiple “success stories” within the genetic genealogy community, such as reunions of adoptees with their biological families (7).

In the last few months, law enforcement agencies have started exploiting third-party consumer genomics services to trace suspects by finding their distant genetic relatives. This route to identify individuals, dubbed long-range familial search, has been predicted before (8) and offers a powerful alternative to familial searches in forensic databases, which can only identify close (1st-2nd degree) relatives (9, 10) and is highly regulated (11). In one notable case, law enforcement used a long-range familial search to trace the Golden State Killer (12, 13). Investigators generated a genome-wide profile of the perpetrator from a crime scene sample and uploaded the profile to GEDmatch ~1 million DNA profiles. The GEDmatch search identified a 3rd-degree cousin (12). Extensive genealogical data traced the identity of the perpetrator, which

was confirmed by a standard DNA test. Between April to August 2018, at least 13 cases were reportedly solved by long range familial searches (Table 1 and table S2). Most of these investigations focused on cold cases, for which decades of investigation failed to identify the offender. Nonetheless, one case involved a crime from April 2018, suggesting that some law enforcement agencies have incorporated long-range familial DNA searches into active investigations. Parabon Nanolabs, a forensic DNA company, have announced that they set up a division that will use long-range familial searches and have already uploaded 100 cold cases to third-party DTC services (14). All of these suggest that long-range familial searches may become a standard investigative tool.

We took an empirical approach to investigate the probability that a long-range familial search will identify an individual. To this end, we analyzed a dataset of 1.28 million individuals who were tested with a DTC provider (15). We retained relatives with at least two IBD segments of >6cM each to increase the chance of correctly inferring genealogical relationships. Next, we removed pairs with IBD segments greater than 700cM (i.e., first cousin and closer relationships) to circumvent ascertainment biases due to the tendency of close relatives to undergo genetic testing together. Finally, considering each individual in turn as our “target”, we counted the number of individuals with a total IBD sharing between 30cM and 600cM with the target (15). The low end of our range corresponds to ~4th cousins and the high end to 2nd cousins, on the basis of a crowd sourcing project (16).

Our results show that nearly 60% of long-range familial searches return a relative with IBD segments with a total length of 100cM or more (Fig. 1A). This level of IBD sharing

usually corresponds to a 3rd cousin or closer relative, similar to the case of the Golden State Killer. Interestingly, these success rates are higher than with surname inference from the Y-chromosome, which is another genetic re-identification tactic (17). In 15% of the searches with our data, the top match had IBD segments of total length at least 300cM, which corresponds to a 2nd cousin or closer relative. We validated our results by performing 30 random long-range familial searches in GEDmatch. The results were similar: the top match in GEDmatch shared >100cM in 76% of the cases (CI: 59%-88%) and >300cM in 10% (CI: 3%-25%) of the searches, similar to the results with our 1.28M million individuals (Fig. 1A).

Long-range familial searches create racial disparity that is the opposite of disparities documented in traditional forensic databases (11). About 75% of the 1.28M individuals were primarily of North European genetic background (fig. S2 and table S3), similar to previous reports of DTC genomics data (18). Individuals of primarily North European background were 30% more likely to have a >100cM match than individuals whose genetic background was primarily from sub-Saharan Africa (fig. S3).

More broadly, a genetic database needs to cover only 2% of the target population to provide a 3rd cousin match to nearly any person (Fig. 1B). This assertion relies on a population genetics model that takes into account the probability of sharing at least two IBD segments of length >6cM and assumes the population growth rates over the last 200 years in the Western world (15) (fig. S4). This model has multiple simplifying assumptions such as no population structure, no inbreeding, and random sampling of participants, and thus should be interpreted only as a rough guideline. Nevertheless, the model showed consistency between our empirical results and the IBD sharing profile of North Europeans in the US (fig. S5). Using the model, we predict that with a database size of ~3 million US individuals of European descent (2% of the adults of this population), over 99% of the people of this ethnicity would have at least a single 3rd cousin match and over 65% are expected to have at least one 2nd cousin match. With the exponential growth of consumer genomics (1), we posit that such database scale is foreseeable for some 3rd party websites in the near future.

Next, we examined the ability to find the person of interest after finding a relative in a long-range familial search. We focused on reducing the search space using basic demographic information, such as geography, age, and sex. Using genealogical records of population-scale family trees (19), we computed the number of relatives of a 3rd cousin match after filtering them on the basis of place of residence, age, and sex. A study of serial criminals indicates that the place of crime is nearly always within 25 miles (20). To be conservative, we thus assumed that the location of the target can be estimated

within a radius of 100 miles. We also assumed that the age of the target can be estimated within a ± 5 yrs interval based on eye witnesses or camera footage, as previously estimated (21). Finally, we assumed the biological sex is known from the DNA sample.

We found that the suspect list can be pruned from basic demographic information. On the basis of counting relevant relatives of the match, the initial list of candidates contains on average ~850 individuals (Fig. 2A). Our simulations indicate that localizing the target to within 100 miles will exclude 57% of the candidates on average (Fig. 2B and table S4). Next, availability of the target's age to within ± 5 yrs will exclude 91% of the remaining candidates (Fig. 2C). Finally, inference of the biological sex of the target will halve the list to just around 16-17 individuals, a search space that is small enough for manual inspection. We also considered a scenario of re-identification of anonymized clinical genetic data. The safe harbor provisions of the HIPAA privacy law permits the release of the year of birth. An age specified at a single year resolution is as expected a more powerful identifier compared to a 10yr interval (Fig. 2D). Together with geography (<100 miles) and sex, it is expected to reduce the search space to just 1-2 individuals (Fig. 2E).

To better understand the risk of re-identifying to human subjects, we conducted a long-range familial search on a specific 1000 Genomes Project individual. We selected a female from the CEU cohort in Utah, whose husband has been identified using surname inference (17). We extracted her genome from the (publicly available) 1000Genomes data repository, re-formatted her genotype to resemble a file released by DTC providers, and uploaded the genotype to GEDmatch. Searching GEDmatch returned two relatives, from North Dakota and Wyoming respectively, with sufficient genetic and genealogical details (Fig. 3). Both relatives shared about 170cM to 180cM with the 1000Genomes sample, which corresponds to 6-7 degrees of separation. They also shared 62cM between each other, indicating that they were distantly related via an ancestral couple who lived 4-6 generations ago. In about one hour of work, we identified the ancestral couple from publicly available genealogical records. Next, we searched for descendants of the ancestral couple that matched the publicly available demographic data of the 1000Genomes sample, such as her expected year of birth and pedigree structure. This step, performed manually, was time consuming and not trivial, as the ancestral couple had over ten children and hundreds of descendants. After a full day of work, we eventually traced the identity of our target, which was the same person we have previously re-identified based on surname inference.

Taken together, we posit that our results warrant a reevaluation of the status quo regarding the identifiability of DNA data, especially of US individuals. While policymakers and the general public may be in favor of such enhanced forensic

capabilities for solving crimes, it relies on databases and services that are open to everyone. Thus, the same technique could also be exploited for harmful purposes, such as re-identification of research subjects from their genetic data. The Revised Common Rule, which will regulate federally funded human subject research starting in January 2019, does not define genome-wide genetic datasets as identifiable information (22). However, the Rule permits the U.S. Department of Health and Human Services to revise the scope of identifiable private information on the basis of technological developments. In light of our results, we encourage HHS to consider genome-wide information as identifiable.

Finally, we propose a measure to mitigate some of the risks and restore control to data custodians. In our proposal, DTC providers should cryptographically sign the text file containing the raw data available to customers (fig. S6). Third-party services will be able to authenticate that a raw genotyping file was created by a valid DTC provider and not further modified. If adopted, our approach has the potential to prevent the exploitation of long-range familial searches to identify research subjects from genomic data. Moreover, it will complicate the ability to conduct unilaterally long-range familial searches from DNA evidence (15). As such, it can complement previous proposals regarding the regulation of long range familial searches by law enforcement (23) and offers better protection in cases where the law cannot deter misuse. To facilitate consideration of our approach by the community, we provide a demo source code on GitHub that can sign and verify the raw genotype files using previously-published digital signature scheme (24). Overall, we believe that technical measures, clear policies for law enforcement in using long-range familial searches, and respecting the autonomy of participants in genetic studies are necessary components for long term sustainability of the genomics ecosystem.

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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/10.1111/j.1556-4029.2011.01867.x) [Medline](#)
10. N. A. Garrison, R. V. Rohlf, S. M. Fullerton, Forensic familial searching: Scientific and social implications. *Nat. Rev. Genet.* **14**, 445 (2013). [doi:10.1038/nrg3519](https://doi.org/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](https://doi.org/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-rape-case/.
13. J. Jouvenal, "To find alleged Golden State Killer, investigators first found his 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-grandparents/2018/04/30/3c865fe7-dfcc-4a0e-b6b2-0bec548d501f_story.html?utm_term=.6ff5cfff1630e.
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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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-rape-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/ramond-dj-freez-rowe-arrested-for-murder-of-schoolteacher-christy-mirack/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-rape-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.

ACKNOWLEDGMENTS

We thank G. Japhet and A. Gordon for their contributions to the cryptographic signature scheme, Y. Naveh, Y. Ben-David, C. Moore, and the DNA Doe Project for valuable comments. **Funding:** Y.E. holds a Burroughs Wellcome Fund Career Awards at the Scientific Interface. S.C. thanks Israel Science Foundation grant

no. 407/17. **Author contributions:** Y.E. conceived the idea for this study. Y.E. and T.S. conducted the analysis of matches using the MyHeritage and the Geni.com data. S.C. and I.P. developed the theoretical framework to estimate the number of matches. Y.E. and S.C. conducted the trace back of the 1000Genomes sample. **Competing interests:** Y.E. and T.S. adapted the code for the cryptographic signatures. Y.E., T.S., I.P. and S.C. wrote the manuscript. Y.E. and T.S. are MyHeritage employees. Y.E. is also a consultant of ArcBio. I.P. holds equity in 23andMe. S.C. is a paid consultant of MyHeritage. When multiple companies are mentioned in this manuscript, we listed them in a lexicographic order. **Data and materials availability:** the code for the cryptographic signatures is available on <https://github.com/erlichya/signature> with an MIT license. The millions of genealogical records for the demographic analysis data are available on <http://familinx.org/> under Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. The code for the population genetics simulation and 1000Genomes extraction is available in the Supplementary Material under an MIT License. Following the MyHeritage Terms, we cannot share the individual level genomic data. We will share the anonymized IBD network topology on request and subject to MyHeritage Terms and Conditions and Privacy Policy under the following terms: (i) researchers will need an IRB approval for their study (ii) the data can only be processed in a MyHeritage facility and cannot be used to re-identify individuals (iii) the results can only be used for non-commercial purpose (iv) MyHeritage does not ask authorship in new publications that uses the anonymized IBD network. Researchers who are interested in the data or to pursue research collaboration opportunities can contact dnaresearch@myheritage.com.

SUPPLEMENTARY MATERIALS

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

Materials and Methods

Figs. S1 to S6

Tables S1 to S4

References (25–44)

14 June 2018; accepted 1 October 2018

Published online 11 October 2018

10.1126/science.aau4832

Fig. 1. The performance of long range familial searches for various database sizes. (A) The probability to find at least one relative for various IBD thresholds (top) with 1.28 million searches of DTC tested individuals (red) and 30 random GEDmatch searches (gray). Light gray: 95% confidence interval for the GEDmatch estimates. Dashed line: the probability of a surname inference from Y chromosome data (21). Bottom: 95% confidence intervals (circles) and average (squares) total IBD length for 1st cousin once removed (1C1R) to 4th cousin once removed (20). (B) A population-genetic theoretical model for the probability to find relatives up to a certain type of cousinship as a function of the database coverage of the population. 1C-4C denote 1st to 4th cousins.

Fig. 2. Tracing a person of interest from a distant match using demographic identifiers. (A) The possible relatives of a match (green) in a database. Each square represents a potential degree of relatedness. The range corresponds to the 5%-95% percentile of shared IBD in cM from ref: (16). Red: relatives that could fit a *bona-fide* 3C match (~100cM). The average number of relatives is denoted in the top-left corner of each square based on a fertility rate of 2.5 children per couple. Nie/Nep: Niece/Nephew; G2: Great-great; G3: Great-great-great; A/U: Aunt/Uncle. (B) An example of the geographical dispersion of 3rd cousins or 2nd cousins once removed around the matched relative. Every circle denotes 100km. (C and D) The distribution of the expected age differences between matches and their potential relatives with a genetic distance of third cousins. The main text reports a conservative scenario, in which the age estimator of the target is in the highest bin of each histogram (red arrow). The age distribution is shown in (C) 10yr resolution in and (D) in 1yr resolution. (E) The entire pipeline of using demographic identifiers along with a long-range familial match to identify a US person (blue: average number of people).

Fig. 3. Tracing a 1000Genomes sample using long range familial search. In black: the CEU pedigree. To respect the privacy of the family, we omitted the sample identifiers and the exact pedigree structure. A GEDmatch search of the person of interest (black circle) returned two males (squares with gray dots) with a total IBD sharing of 180cM and 171cM to the target, respectively, and 62cM between themselves. Using public genealogical records, we identified the ancestral couple (asterisk) of the matches and the person of interest.

Table 1: Public cases of long range familial cases.

Case	Announcement	By	Closest match	Comments
Buckskin Girl	April 9, 2018	DNA Doe Project	First cousin once removed	
Golden State Killer	April 24, 2018	Barbara Rae-Venter	Third cousin	
Lyle Stevik	May 8, 2018	DNA Doe Project	Second cousin	Inbreeding complicated the estimation of the match.
William Earl Talbott II	May 21, 2018	Parabon	Half first cousin once removed	Second cousins were identified as well
Joseph Newton Chandler III	June 21 2018	DNA Doe Project	Second cousin once removed	
Gary Hartman	June 22, 2018	Parabon	Half first-cousin	Genealogists were able to overcome non-paternity event in the tree of the suspect.
Raymond "DJ Freez" Rowe	June 25, 2018	Parabon	-	
James Otto Earhart	June 26, 2018	Parabon	Second cousin	
John D. Miller	July 15, 2018	Parabon	-	
Matthew Dusseault/Tyler Grenon	July 28, 2018	Parabon	-	
Spencer Glen Monnett	July 29, 2018	Parabon	-	Active case for a crime that occurred in April 2018
Darold Wayne Bowden	August 23, 2018	Parabon	-	
Michael F. Henslick	August 29, 2018	Parabon	-	