# Identifying Personal Genomes by Surname Inference

Melissa Gymrek,[1,2,3,4] Amy L. McGuire,[5] David Golan,[6] Eran Halperin,[7,8,9] Yaniv Erlich[1]*

Sharing sequencing data sets without identifiers has become a common practice in genomics. Here, we report that surnames can be recovered from personal genomes by profiling short tandem repeats on the Y chromosome (Y-STRs) and querying recreational genetic genealogy databases. We show that a combination of a surname with other types of metadata, such as age and state, can be used to triangulate the identity of the target. A key feature of this technique is that it entirely relies on free, publicly accessible Internet resources. We quantitatively analyze the probability of identification for U.S. males. We further demonstrate the feasibility of this technique by tracing back with high probability the identities of multiple participants in public sequencing projects.

Surnames are paternally inherited in most human societies, resulting in their co-segregation with Y-chromosome haplotypes (*1*–*5*). Based on this observation, multiple genetic genealogy companies offer services to reunite distant patrilineal relatives by genotyping a few dozen highly polymorphic short tandem repeats across the Y chromosome (Y-STRs). The association between surnames and haplotypes can be confounded by nonpaternity events, mutations, and adoption of the same surname by multiple founders (*5*). The genetic genealogy community addresses these barriers with massive databases that list the test results of Y-STR haplotypes along with their corresponding surnames. Currently, there are at least eight databases and numerous surname project Web sites that collectively contain hundreds of thousands of surname-haplotype records (table S1).

The ability of genetic genealogy databases to breach anonymity has been demonstrated in the past. In a number of public cases, male adoptees and descendants of anonymous sperm donors used recreational genetic genealogy services to genotype their Y-chromosome haplotypes and to search the companies' databases (*6*–*9*). The genetic matches identified distant patrilineal relatives and pointed to the potential surnames of their biological fathers. By combining other pieces of demographic information, such as date and place of birth, they fully exposed the identity of their biological fathers. Lunshof *et al*. (*10*) were the first to speculate that this technique could expose the full identity of participants in sequencing projects. Gitschier (*11*) empirically approached this hypothesis by testing 30 Y-STR haplotypes of CEU participants in these databases and reported that potential surnames can be detected. [CEU participants are multigenerational families of northern and western European ancestry in Utah who had originally had their samples collected by CEPH (Centre d'Etude du Polymorphisme Humain) and were later reconsented to participate in the HapMap project.] However, these surnames could match thousands of individuals, and the study did not pursue full re-identification at a single-person resolution.

Our goal was to quantitatively approach the question of how readily surname inference might be possible in a more general population, apply this approach to personal genome data sets, and demonstrate end-to-end identification of individuals with only public information. We show that full identities of personal genomes can be exposed via surname inference from recreational genetic genealogy databases followed by Internet searches. In all cases in which individuals were studied who had donated DNA samples, the informed consent statements they had signed stated privacy breach as a potential risk and the data usage terms did not prevent re-identification. Representatives of relevant organizations that funded the original studies were notified and confirmed the compliance of this study with their guidelines (*12*).

As a primary resource for surname inference, we focused on Ysearch (www.ysearch.org) and

[1]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142, USA. [2]Harvard–Massachusetts Institute of Technology (MIT) Division of Health Sciences and Technology, MIT, Cambridge, MA 02139, USA. [3]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [4]Department of Molecular Biology and Diabetes Unit, Massachusetts General Hospital, Boston, MA 02114, USA. [5]Center for Medical Ethics and Health Policy, Baylor College of Medicine, Houston, TX 77030, USA. [6]Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv 69978, Israel. [7]School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel. [8]Department of Molecular Microbiology and Biotechnology, Tel-Aviv University, Tel Aviv 69978, Israel. [9]The International Computer Science Institute, Berkeley, CA 94704, USA.

*To whom correspondence should be addressed. E-mail: yaniv@wi.mit.edu

SMGF (www.smgf.org), the two largest public genetic genealogy databases with free-of-charge, built-in search engines. The interfaces of these engines are quite similar and allow users to insert a combination of Y-STR alleles and search for matching records on the basis of genetic similarity. The retrieved records contain surnames typically with information about the patrilineal line, such as geographical locations, potential spelling variants, and pedigrees. In total, these databases contain ~39,000 unique surname entries from ~135,000 records. The distribution of records per surname is significantly correlated ($R^2 = 0.78$, $P < 1.20 \times 10^{-6}$) with surname frequencies in the United States, suggesting an overall good representation of this population (Fig. 1A).

To test the probability of surname inference, we challenged the two databases with an orthogonal cohort of Y-STR haplotypes consisting of 34 markers (table S2) from 911 individuals, primarily with Caucasian ancestry, whose surnames are known (table S3). This cohort was compiled from YBase, a distinct genetic genealogy database, and contains individuals with 521 surnames that segregate in the U.S. population. In each haplotype query, our surname recovery algorithm began by retrieving the database record with the shortest time to most recent common ancestor (TMRCA) with the input haplotype (fig. S1 and table S4). Then, it calculated a confidence score that the surname match of the retrieved record is significantly better than other matches. If the score passed a user-defined threshold, the algorithm assigned the record's surname to the input haplotype; otherwise, it categorized it as "unknown." We tested the algorithm with a range of confidence thresholds to explore the trade-off between successful versus wrong recovery of surnames. Finally, we weighted the results using a stratified sampling approach to reflect the frequency of surnames in the U.S. population (*13*).

Our analysis projects a success rate of ~12% (SD = 2%) in recovering surnames of U.S. Caucasian males (Fig. 1B and fig. S2). This rate can be accomplished with a conservative threshold that would return a wrong surname in 5% of cases and label 83% of cases as unknown. Higher success rates of up to 18% can be achieved at the price of increased probability to recover an incorrect surname. Because our input cohort is based on individuals who were tested with genetic genealogy services, our results are presumably mostly relevant to socio-economic groups with high participation in these services—namely, upper- and middle-class U.S. Caucasians.

Combining the recovered surname with additional demographic data can narrow down the identity of the sample originator to just a few individuals. The analysis above indicated that most recovered surnames are quite rare, with frequencies of less than 1:4000 in the U.S. population, corresponding to <40,000 males (Fig. 1C and fig. S3) (*13*). We considered a scenario in which the genomic data are available with the target's year of birth and state of residency, two identifiers that are not protected by the United States Health Insurance Portability and Accountability Act (HIPAA) (*14*). Searching individuals by year of birth, state, and surname combinations is supported by various online public record search engines, such as PeopleFinders.com or USA-people-search.com. On the basis of extensive simulations with the U.S. Census data, our results predict that year of birth and state alone are weak identifiers and searches based on their combination would match at least 60,000 U.S. males in 50% of cases (Fig. 1D). However, when surname information is added to the search, the median list size shrinks to only 12 males, which are few enough matches to investigate individually.

Next, we established the feasibility of Illumina sequencing to produce accurate Y-STR haplotypes. Using lobSTR, an algorithm for STR

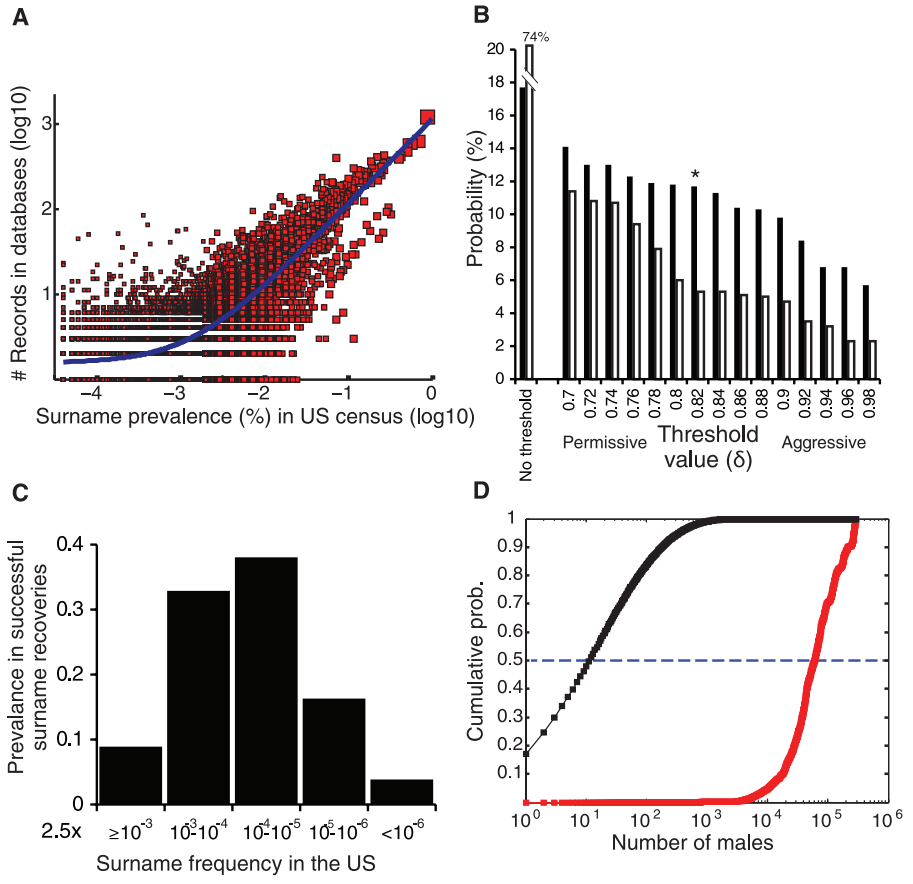


**Fig. 1.** Quantitative assessment of identification via surname inference. (**A**) The number of Ysearch and SMGF records as a function of surname prevalence in the U.S. population. The best-fit line is shown in blue. (**B**) Expected performance of surname recovery. The probability of successful recovery (closed bars) and wrong recovery (open bars) is shown at different surname confidence thresholds. The star indicates the middle-range performance threshold that was described in the main text. (**C**) The expected distribution of recovered surnames as a function of their prevalence. Most recovered surnames are expected to have a frequency of 1:4000 individuals or less. (**D**) The cumulative distribution function of U.S. males with a profile that matches a specific age, state, and surname combination (black) compared to the distribution when only age and state are known (red). The median is labeled with a dashed line.

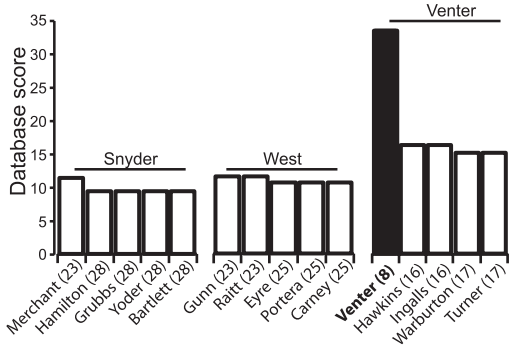


**Fig. 2.** The top five records retrieved after searching Ysearch with the Y-STR haplotypes of Michael Snyder, John West, and Craig Venter. The expected number of generations to the MRCA is given in parentheses for each record. Searching with Craig Venter returned a "Venter" record (closed bar) as the top match.

**Table 1.** Comparison of CEU identification cases.

| Feature | Pedigree 1 | | Pedigree 2 | | Pedigree 3 |
|---|---|---|---|---|---|
| Genome for surname recovery | Paternal grandfather | Maternal grandfather | Paternal grandfather | Maternal grandfather | Father |
| Surname freq. in U.S.* | Rare | Rare | Common | Rare | Rare |
| Meioses between target and source | 3 | 5 | 5 | 7 | 2 |
| Relationship between target and source | Nephew | First cousin once removed | Great-great nephew | Second cousin once removed | Grandchild |
| Supporting evidence | State of residency, pedigree structure, age, and maiden name are the same | | State of residency, pedigree structure, age, and maiden name are the same | | State of residency, pedigree structure are the same (ages are not given) |
| $P$ (random match)† | $<5 \times 10^{-9}$ | | $<5 \times 10^{-6}$ | | $<10^{-5}$ |

*Common: surnames with a prevalence of $>10^{-4}$; Rare: surnames with a prevalence of $\leq 10^{-4}$. †The estimated probability of finding at least one family with the same characteristics after scanning all Utah households.

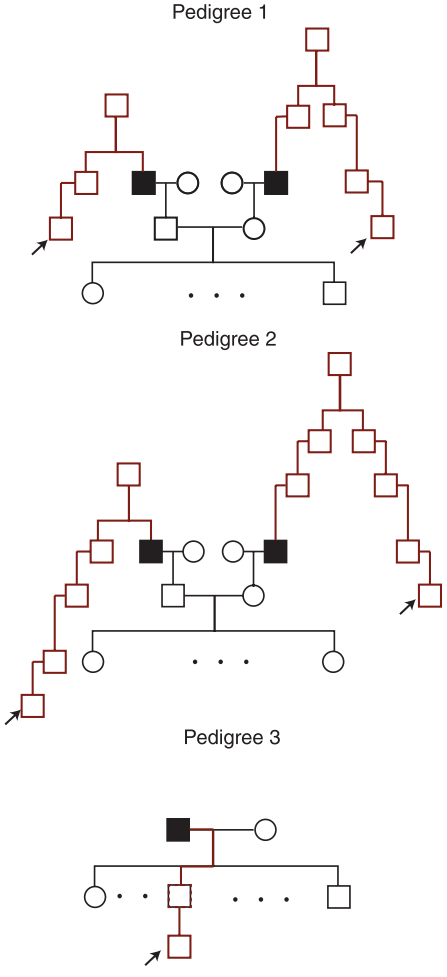


**Fig. 3.** Illustrations of the three CEU pedigrees (black) showing how genetic information from distant patrilineal relatives (arrow; red, patrilineal lines) can identify individuals. Filled squares represent sequenced individuals. To respect the privacy of these families, only abbreviated versions are presented. The sex of the CEU grandchildren was randomized. The numbers of grandchildren are not given.

profiling from raw sequencing reads (*15*), we processed 10 high-coverage male genomes from the Human Genome Diversity Panel (HGDP). lobSTR produced Y-STR haplotypes with an average number of 53 out of the possible 79 genealogical markers (table S5). Comparing these haplotypes to capillary electrophoresis results revealed 99% accuracy. We further found that even at lower sequencing coverage of 10×, informative haplotypes can be obtained by lobSTR (fig. S4). To test the ability to retrieve genetic genealogy records with the Illumina haplotypes, we profiled STRs from the genome of a U.S. Caucasian male from our lab collection that was sequenced with Illumina 100–base pair (bp) reads to a coverage of 13×. In parallel, we submitted this sample to the genealogy service of Sorenson Genomics and created a Ysearch record based on their results. A search with the Illumina haplotype returned his Ysearch entry as a top record (fig. S5).

The National Center for Biotechnology Information archives host a small number of genomes from identified individuals, providing good test cases for identification via surname inference. We used lobSTR to extract Y-STR haplotypes from the genomes of John West (*16*), Michael Snyder (*17*), and Craig Venter (*18*) (table S6). Searching Ysearch and SMGF with the Y-STR haplotypes of West and Snyder did not return their surnames and resulted in low matches to records with relatively ancient MRCAs 23 to 28 generations ago (*13*). A search with Craig Venter's haplotype returned a clear match to a "Venter" record that was concordant at all 33 comparable markers and with an estimated TMRCA of less than eight generations (Fig. 2 and table S7). We further tested whether it would be feasible to trace back Craig Venter by combining the inferred surname with demographic profiling. A query for "Surname: Venter; Year of Birth: 1946; State: California" in online public record search engines retrieved two matching records of males, one of whom was Craig Venter himself.

Surname inference from personal genomes puts the privacy of current de-identified public data sets at risk (*19*). We focused on the male genomes in the collection of Utah Residents with Northern and Western European Ancestry (CEU). The informed consent of these individuals did not definitively guarantee their privacy and stated that future techniques might be able to identify them (*20*). To test the ability to trace back the identities of these samples from personal genomes, we processed with lobSTR 32 Illumina genomes of CEU male founders that reside in public repositories of the 1000 Genomes Project (*21*) and the European Nucleotide Archive that were sequenced with read lengths of at least 76 bp. Most of these genomes were sequenced to a shallow depth of less than 5× and produced sparse Y-STR haplotypes. We selected the 10 genomes that had the most complete Y-STR haplotypes with a range of 34 to 68 markers to attempt surname recovery. Searching the genetic genealogy databases returned top-matching records with Mormon ancestry in 8 of the 10 individuals for whom the top hit had at least 12 comparable markers. Moreover, for four individuals, the top match consisted of multiple records with the same surname, increasing the confidence that the correct surname was retrieved. This potentially high surname recovery rate stems from a combination of the deep interest in genetic genealogy among this population and the large family sizes, which exponentially increases the number of targeted individuals for every person who is tested.

In five surname recovery cases, we fully identified the CEU individuals and their entire families with very high probabilities (Table 1). These five cases belonged to three pedigrees, in two of which the surnames of both the paternal and maternal grandfathers were recovered. Our strategy for tracing back individuals relied on the recovered surnames as well as publicly available Internet resources such as record search engines, obituaries, and genealogical Web sites, and demographic metadata available in the Coriell Cell Repository Web site. The year of birth was inferred by subtracting the ages in Coriell from the year of collecting samples. Each complete pedigree re-identification took 3 to 7 hours by a single person. The identified families matched exactly to the corresponding pedigree descriptions in the Coriell database: The number of children, the birth order of daughters and sons, and the state of residence were identical. All grandparents were alive in 1984, the year that the CEU cell line collection was established (*22*). In the two cases of a dual surname recovery from both grandfathers, the surname of the father

and the maiden name of the mother matched exactly to the grandfathers' surnames, substantially increasing the confidence of the recovery. Coriell also lists the ages (*23*) during sample collection for these two pedigrees, which agreed with the age differences of all tested cases with the identified family members. Using genealogical Web sites, we traced the patrilineal lineage that connects each identified genome through the MRCA to the record originator in the genetic genealogy database (Fig. 3). This analysis revealed that two to seven meiosis events link the CEU genome to the record source. Finally, we calculated that the probability of finding random families in the Utah population with these exact demographic characteristics is less than 1 in $10^5$ to $5 \times 10^9$ (*13*). In total, surname inference breached the privacy of nearly 50 individuals from these three pedigrees.

This study shows that data release, even of a few markers, from one person can spread through deep genealogical ties and lead to the identification of another person who might have no acquaintance with the person who released his genetic data. The propagation of information through shared male lines amplifies the range of identification, allowing ~135,000 records to potentially target several million U.S. males. Another feature of this identification technique is that it entirely relies on free, publicly available resources. It can be completed end-to-end with only computational tools and an Internet connection. The compatibility of our technique with public record search engines makes it much easier to continue identifying other data sets in the same pedigree, including female genomes, once one male target is identified. We envision that the risk of surname inference will grow in the future. Genetic genealogy enthusiasts add thousands of records to these databases every month. In addition, the advent of third-generation sequencing platforms with longer reads will enable even higher coverage of Y-STR markers, further strengthening the ability to link haplotypes and surnames.

Similar to other genetic privacy issues (*24*–*30*), preventing surname inference from public whole-genome data sets might be quite challenging. Masking Y-STR markers could limit the effectiveness of the method presented in this study, but this approach is not sustainable (*13*). Our analysis suggests that Y-STR haplotypes can be imputed back from single-nucleotide polymorphisms (SNPs) on the Y chromosome (Y-SNPs) when a large reference set of male genomes will be available (fig. S6). In addition, community efforts, such as the Y Chromosome Genome Comparison, have already started exploring the association between Y-SNPs and surnames (table S1) and might allow bypassing Y-STR masking. We also posit that restricting genetic genealogy information is not practical, as some of the data are already scattered in multiple end-user Web sites and genealogy mailing lists.

Existing policy tools, such as controlled-access databases with data use agreements, may mediate the exposure of genomic information to surname inference. However, in our view, the appropriate response to genetic privacy challenges is not for the public to stop donating samples or for data sharing to stop. These would be devastating reactions that could substantially hamper scientific progress. Rather, we believe that establishing clear policies for data sharing, educating participants about the benefits and risks of genetic studies (*31*), and the legislation of proper usage of genetic information (*32*) are pivotal ingredients to support the genomic endeavor.

## References and Notes


1. B. Sykes, C. Irven, *Am. J. Hum. Genet.* **66**, 1417 (2000).
2. T. E. King, S. J. Ballereau, K. E. Schürer, M. A. Jobling, *Curr. Biol.* **16**, 384 (2006).
3. B. McEvoy, D. G. Bradley, *Hum. Genet.* **119**, 212 (2006).
4. T. E. King, M. A. Jobling, *Mol. Biol. Evol.* **26**, 1093 (2009).
5. T. E. King, M. A. Jobling, *Trends Genet.* **25**, 351 (2009).
6. R. Lehmann-Haupt, "Are sperm donors really anonymous anymore?" *Slate*, 1 March 2010.
7. G. Naik, "Family secrets: An adopted man's 26-year quest for his father," *Wall Street Journal*, 2 May 2009.
8. R. Stein, "Found on the Web, with DNA: A boy's father," *Washington Post*, 13 November 2005.
9. A. Motluk, *New Sci.* **188**, 6 (3 November 2005).
10. J. E. Lunshof, R. Chadwick, D. B. Vorhaus, G. M. Church, *Nat. Rev. Genet.* **9**, 406 (2008).
11. J. Gitschier, *Am. J. Hum. Genet.* **84**, 251 (2009).
12. L. L. Rodriguez, L. D. Brooks, J. H. Greenberg, E. D. Green, *Science* **339**, 275 (2013).
13. See supplementary materials on *Science* Online.
14. *Federal Register: 45 CFR 164.514(b-c)* (2002).
15. M. Gymrek, D. Golan, S. Rosset, Y. Erlich, *Genome Res.* **22**, 1154 (2012).
16. N. Leat, L. Ehrenreich, M. Benjeddou, K. Cloete, S. Davison, *Forensic Sci. Int.* **168**, 154 (2007).
17. S. K. Lim, Y. Xue, E. J. Parkin, C. Tyler-Smith, *Int. J. Legal Med.* **121**, 124 (2007).
18. S. Levy *et al.*, *PLoS Biol.* **5**, e254 (2007).
19. Full details of this analysis were provided to the reviewers. However, they are not presented here to respect the privacy of these families. Further inquiries can be made to the corresponding author.
20. http://hapmap.ncbi.nlm.nih.gov/downloads/elsi/CEPH_Reconsent_Form.pdf.
21. The 1000 Genomes Project Consortium, *Nature* **467**, 1061 (2010).
22. S. M. Prescott, J. M. Lalouel, M. Leppert, *Annu. Rev. Genomics Hum. Genet.* **9**, 347 (2008).
23. Based on the results of this study, the NIH removed the ages from Coriell to a secure location (*12*).
24. Z. Lin, A. B. Owen, R. B. Altman, *Science* **305**, 183 (2004).
25. F. R. Bieber, C. H. Brenner, D. Lazer, *Science* **312**, 1315 (2006).
26. N. Homer *et al.*, *PLoS Genet.* **4**, e1000167 (2008).
27. K. B. Jacobs *et al.*, *Nat. Genet.* **41**, 1253 (2009).
28. H. K. Im, E. R. Gamazon, D. L. Nicolae, N. J. Cox, *Am. J. Hum. Genet.* **90**, 591 (2012).
29. D. W. Craig *et al.*, *Nat. Rev. Genet.* **12**, 730 (2011).
30. E. E. Schadt, S. Woo, K. Hao, *Nat. Genet.* **44**, 603 (2012).
31. A. L. McGuire, R. A. Gibbs, *Science* **312**, 370 (2006).
32. Presidential Commission for the Study of Bioethical Issues, Privacy and Progress in Whole Genome Sequencing. Privacy and Progress in Whole Genome Sequencing (2012).


**Acknowledgments:** We thank FamilyTreeDNA and SMGF for technical assistance. We also thank D. Esposito, A. Goren, G. Fink, D. Page, W. Kramer, and R. Ronen for useful discussions. Y.E. is an Andria and Paul Heafy Family Fellow. This publication was supported by a gift from Jim and Cathy Stone (Y.E.), the National Defense Science and Engineering Graduate Fellowship (M.G.), and the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University (D.G. and E.H.).

## Supplementary Materials

www.sciencemag.org/cgi/content/full/339/6117/321/DC1
Supplementary Text
Figs. S1 to S6
Tables S1 to S7
References

31 August 2012; accepted 3 December 2012
10.1126/science.1229566