HUMAN GENETICS

Deriving genomic diagnoses without revealing patient genomes

Karthik A. Jagadeesh,[1]* David J. Wu,[1]* Johannes A. Birgmeier,[1] Dan Boneh,[1,2]† Gill Bejerano[1,3,4]†

Patient genomes are interpretable only in the context of other genomes; however, genome sharing enables discrimination. Thousands of monogenic diseases have yielded definitive genomic diagnoses and potential gene therapy targets. Here we show how to provide such diagnoses while preserving participant privacy through the use of secure multiparty computation. In multiple real scenarios (small patient cohorts, trio analysis, two-hospital collaboration), we used our methods to identify the causal variant and discover previously unrecognized disease genes and variants while keeping up to 99.7% of all participants' most sensitive genomic information private.

Personalized genomics is transforming 21st-century medicine. It has opened a window into human cancer genomics and has provided new insights into the combinatorics fueling complex polygenic diseases. Personalized medicine's first major triumph lies in the field of monogenic diseases, estimated to affect up to 10% of individuals. Personalized genomics now offers a definitive diagnosis for more than 4500 monogenic diseases and may soon offer diagnosis for all 7000 (*1*). In comparison, fewer than 50 digenic diseases have been identified (*2*), and our understanding of more complex diseases is even more limited (*3*). Frequency-based filters have proven extremely effective in diagnosing monogenic diseases (*4*). In essence, variants found in a control population (common variants) are likely benign (*5*), whereas functional rare variants not found in the control population but seen in multiple affected individuals are likely to cause disease (*6*–*8*). These filters seek the gene or variant present in all (or most) affected individuals but in none (or very few) of the unaffected individuals.

For example, one can take a small cohort of unrelated individuals suspected of suffering from the same genetic disorder and compare their genomes to those of tens of thousands of unaffected individuals [e.g., from the Exome Aggregation Consortium (ExAC) (*5*)]. As we show below, in many scenarios, the gene with rare functional mutations in most patients in a small cohort is responsible for their condition.

Frequency-based computation highlights the fundamental "serve or protect" dilemma of all genomic data. To find the root cause of a patient's disease, one wishes to compare the patient's genome to as many other genomes as possible, both affected and unaffected, related and unrelated. To advance modern medicine, all sequenced genomes should be shared. However, one's genome continues to reveal more and more about oneself, including susceptibility to a variety of diseases. Individuals with genetically associated disease phenotypes will be particularly resistant to sharing such information, fearing discrimination and bias (*9*). To protect its owner and next of kin, no sequenced genome should be shared.

To date, this dilemma has been solved by allowing institutions unrestricted access to all the genomes in their possession. Limited sharing between institutions is done by providing obfuscated summary statistics (*10*). Current methods for sharing genomes have shortcomings that make them suboptimal. Providing full access at individual institutions allows for too much information to be shared in certain situations (*11*). Disease-specific "beacons" (that is, web servers that answer allele-presence queries) are prone to attacks that can identify individuals participating in the study (*12*). Beacons also only provide allele-presence query capabilities and do not have the flexibility needed for analyzing variant interactions within an individual (*13*). Additionally, it is risky to share genomic data with third-party services specializing in genomic analysis, as these services may wish to further monetize these data.

We introduce a proof-of-concept cryptographic implementation that both serves and protects. Although many millions of genomic variants from all individuals are needed to perform the computation, only a handful of causal variants are ultimately of interest for the purpose of a diagnosis (e.g., just the rare variants in the single gene that is mutated across many patients).

We take a two-step approach: First, we convert patient genomes into vectors of simple values and show how simple operations on these vectors reveal the causative variant(s) (fig. S1). Then, we apply a cryptographic method called Yao's protocol to perform the desired computation without revealing any participant's input (fig. S2). To illustrate this, imagine Alice and Bob each hold a secret number between 1 and 10. Alice and Bob want to determine whether Bob's number is larger than Alice's without revealing their numbers to each other. There are $10 \times 10 = 100$ possible combinations of values between Alice and Bob. In 45 of them, Bob's number is larger, and in 55, it is not. Alice prepares 100 notes with every possible outcome: 45 say "Bob's number is larger," and 55 say "Bob's number isn't larger." Alice then constructs two sets of 10 different keys each: one set of 10 keys corresponding to each value she may hold and another set corresponding to each value Bob may hold. For each of the 100 combinations of values held by Alice and Bob, Alice places the corresponding note in one box and double-locks this box with two keys: one for her possible value and one for Bob's possible value. For example, she places a note saying "Bob's number isn't larger" in a box and double-locks it with a key from the first set labeled "Alice holds the number 9" and a key from the second set labeled "Bob holds the number 9." Alice places Bob's 10 labeled keys on a table and leaves the room. Bob enters the room and picks up the key for the value he holds. He removes the label from his chosen key and goes to a second room where Alice has left him all 100 unlabeled double-locked boxes (in a random order) and one unlabeled key that matches the number she holds. Bob tries to open each of the 100 doubly locked boxes using the two keys he has. By design, only one box will open, and that box contains the note saying whether Bob holds a larger number or not.

To apply Yao's protocol to genomic diagnosis, we assume that each individual involved in a study has private access to their exome (or genome). If we are looking to identify a causal variant, we provide each individual a variant vector of all possible rare missense and nonsense variants in the human genome (of length 28,413,589 bases for the human exome). We ask them to privately denote (using simple code) "true" or "false" next to each variant (to indicate whether they have the specific mutation or not, respectively). If we are looking to identify a causal gene, we provide each individual with a gene vector of 20,663 genes in the human genome from *A1BG* to *ZZZ3*. We ask them to write "1" next to a gene if they have one or more rare functional variants in this gene; otherwise, they are instructed to write "0," using very simple code on their own computers (fig. S1, A and B).

We define three simple Boolean operations (INTERSECTION, SETDIFF, and MAX) that are useful for patient diagnosis (fig. S1C). INTERSECTION of two variant vectors reveals all rare functional variants that two parties share. SETDIFF of an affected and unaffected individual's variant vector allows us to discard variants seen in healthy individuals. The maximum (MAX) operation can be used to find a gene containing rare functional mutations in the greatest number of affected cases (fig. S1C). Although Yao's protocol provides a simple and efficient solution for secure

[1]Department of Computer Science, Stanford University, Stanford, CA 94305, USA. [2]Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA. [3]Department of Developmental Biology, Stanford University, Stanford, CA 94305, USA. [4]Division of Medical Genetics, Department of Pediatrics, Stanford University, Stanford, CA 94305, USA.
*These authors contributed equally to this work. †Corresponding author. Email: dabo@cs.stanford.edu (D.B.); bejerano@stanford.edu (G.B.)

two-party computation, in many of the scenarios we describe, the computation occurs among multiple parties (e.g., many individuals, each with their personal genomes). Here, we reduce the general problem of secure multiparty computation to that of secure two-party computation by working in a "two-cloud" model (see the supplementary materials and methods section and fig. S2H).

To quantify the privacy guarantees provided by our system, we define the "protection quotient" of a computation to be the fraction of private information that is exposed neither to the other participants nor to the entity running the computation. In our protocols, the protection quotient is the ratio of the total number of patient variants withheld from the output and the total number of patient variants input into the computation. Standard unprotected patient diagnosis operations have a protection quotient of 0% because all values must be exposed to perform the computation.

To demonstrate the utility of our approach, we show three different secure operations over actual patients with causal Mendelian variants (Table 1). Our experiments operate under the two-cloud model, and we place the two clouds on opposite coasts of the United States (see the supplementary materials and methods section and fig. S2H). We use a single-threaded execution for all benchmarks.

The MAX operator can identify the causal gene in small patient cohorts. In our experiments, we used four small cohorts of unrelated individuals suffering from very different rare diseases: Freeman-Sheldon syndrome (FSS), Hadju-Cheney syndrome (HCS), Kabuki syndrome (KaS), and Miller syndrome (MiS). Each individual held a private list of 211 to 374 rare functional variants in 210 to 356 genes (total 767 to 2754 variants per computation). We used the secure MAX function to reveal only the top gene mutated across patients in each cohort. In all four cohorts, we found that the gene mutated in most individuals was the one that has been proven to be the causal gene: *MYH3* in FSS (*6*), *NOTCH2* in HCS (*14*), *KMT2D* in KaS (*8*), and *DHODH* in MiS (*7*) (Table 1, MAX).

In each scenario, our protocol revealed only the variants in the most mutated gene in each cohort while protecting the remaining variants (764 variants in FSS, 1845 in HCS, 2746 in KaS, and 1055 in MiS). This computation has a protection quotient of 99.2 to 99.7% for all four cohort disease data sets. The computation was performed over all 20,663 genes and completed in just 5 to 10 s (Table 1, MAX). The total protocol execution time and bandwidth grow modestly (logarithmically) with the number of cohort individuals participating in the secure computation (fig. S3A).

The SETDIFF operator identifies the causal variant in a trio. We analyzed a trio of an unaffected mother and father and an affected male child with female external genitalia; each holds a list of 164 to 185 (total 524) rare functional variants found in their exomes. The secure SETDIFF operation revealed to the family and test providers two rare variants found in the child but in neither parent (Table 1, SETDIFF). Literature review provided a diagnosis based on one of these two variants: ACTB:p.P164S (with frequency 0 in data from both ExAC and the 1000 Genomes Project), a previously unidentified pathogenic variant of the *ACTB* gene (*15*).

Secure computation kept 522 variants private while revealing only two variants to the test provider and participants. This computation has a protection quotient of 99.6%. The total computation time is 57 min (Table 1, SETDIFF). However, the variant list can easily be split across a small computer array; a typical 30-node cluster reduces the computation time to less than 2 min. The protocol execution time and bandwidth grow logarithmically with the number of family members involved in the secure computation (fig. S3B) and linearly with the size of the variant vector.

The INTERSECTION operator allows two or more genome centers to compare their patient lists to see if they can identify multiple patients with the same rare functional mutation and similar phenotypes while revealing nothing else to each other. For example, we took 928 Washington Mendelian Center (WMC) patients and 282 Baylor Hopkins Center (BHC) patients. For each hospital, we prepared a list of more than 5000 rare functional variants seen in one or more of its patients. Using the secure AND function, the two hospitals found a list of just 159 variants present in both hospitals, pointing at patients who would benefit from phenotype comparison. This short list includes "positive controls" such as known disease variant NOTCH1:p.E694K, which is associated with partial or incomplete penetrance of aortic valve disease (*16*). The WMC and BHC patients were phenotypically characterized with left ventricular outflow defects and thoracic aortic aneurysms, respectively. Examination of the list also suggested previously unrecognized gene-disease associations, such as rare functional variant HCN3:p.R648H (with frequency 5.47×10^{-5} and 0 in data from ExAC and the 1000 Genomes Project, respectively). *HCN3* is a voltage-gated cation channel gene whose mouse knockout causes abnormal ventricular action potential waveform (*17*). In patients from WMC and BHC, this mutation was correlated with dilated cardiomyopathy and coarctation of the aorta, respectively.

Secure computation only revealed 2×159 potential causative variants while protecting the remaining 10,749 variants with a protection quotient of 97.1%. This computation was performed over all rare functional variants in the exome, with a total protocol execution time of 9.4 min (Table 1, INTERSECTION). Because every variant was evaluated independently, the total time and bandwidth scaled linearly with the size of the variant list (fig. S3C). If we wanted to compare not just the exome, occupying 1% of the genome, but the 10% of the genome evolving under purifying selection (*18*), a 30-node compute cluster could be used to reduce the execution time to less than 4 min.

The scenarios we present helped to diagnose real patients and to discover previously unknown gene-disease associations. Complete strangers in disease cohorts (Table 1, MAX) learn nothing about each other except their shared disease-causing gene. For participants for whom the assay does not provide an answer, nothing is revealed. Even for a young nuclear family (e.g., a trio; Table 1, SETDIFF), the test provider (a possible source of discriminatory information) learns almost nothing except the likely disease-causing mutation in the offspring. Moreover, the provider learns virtually nothing about the parents themselves. In the two-hospital scenario (Table 1, INTERSECTION), in which clinicians are incentivized to be the first to diagnose patient conditions, only variants that are worthwhile to compare for both hospitals are revealed, whereas the vast majority of variants remain private to each institute's researchers and patients.

This proof-of-concept work assumes that the protocol participants are incentivized to honestly follow the protocol, but at the end of the protocol execution, they may try to learn some additional information (about other parties' inputs) based on the messages they receive during the protocol execution. We say that a protocol is secure if the only information any party learns by participating in the protocol can be inferred just from that party's input and the overall output of the computation. No current genomic resource can provide this guarantee (*19*, *20*).

Yao's protocol has previously been used in the context of sequence alignment (*21*) and organ transplant compatibility searching (*22*) but not for rare disease diagnosis. Yao's protocol provides an efficient solution for secure two-party computation in the presence of semi-honest adversaries (*23*). There are well-established ways to extend Yao's protocol to provide security against malicious parties who deviate from the protocol description (*24*). Protecting against participants that submit malicious (or malformed) inputs can be done by ensuring that if a participant's variant vector does not meet certain criteria or is not accompanied by an appropriate certificate, then the computation aborts and does not produce any output. Furthermore, the operation-specific "protection quotient" we introduce to assess the fraction of information secured by the computation can also be used to restrict the output returned to all parties if the defined privacy requirements are not met. This approach differs, for example, from differential privacy (*25*), which adds random genomic variation as noise into aggregated summary statistics to avoid individual identification in pooled genomic data (*26*).

The feasibility of a garbled circuit implementation depends on the complexity of the computation. Our benchmarks show that Mendelian computation is feasible even if we extend our variant list 10-fold to accommodate whole-genome (noncoding) analysis (fig. S3). However, some computations, such as those underlying genome-wide association studies (GWASs) with tens of thousands of individuals, are currently impractical to implement using Yao's protocol but may be addressable using technological solutions such as Intel's Software Guard Extensions (*27*). The real challenge in most areas of genomic research

Table 1. Summary of results for different secure genomic multiparty computation scenarios, all using real patient data. N/R, not relevant.

Operation	Relevant information for each operation							Running time measurements		
MAX (over genes)	Scenario: small disease cohort	No. of unrelated probands (who avoid openly sharing their data)	No. of rare functional variants (genes) per proband (median)	No. of probands with rare functional variant/s in gene (top three, descending order)	Gene name	Proven causal gene for disease	Protection quotient (1 – no. of variants shared of top gene/total no. of variants)	Bandwidth (gigabytes)	Compute (s)	Network (s)
	Freeman-Sheldon syndrome	3	258 (253)	3 2 1	*MYH3* *DBT* *ACADVL*	*MYH3*	1 – 3/767 = 99.6%	0.02	0.15	4.91
	Hajdu-Cheney syndrome	7	278 (272)	6 3 3	*NOTCH2* *HLA-DRBI* *MCC*	*NOTCH2*	1 – 6/1853 = 99.7%	0.03	0.18	7.29
	Kabuki syndrome	10	262 (257)	8 3 3	*KMT2D* *COL6A1* *FLNB*	*KMT2D*	1 – 8/2754 = 99.7%	0.04	0.22	9.59
	Miller syndrome	4	267 (258)	4 3 2	*DHODH* *DNAH5* *ACOX2*	*DHODH*	1 – 8/1063 = 99.2%	0.03	0.18	7.29
SETDIFF (over variants)	Scenario: familial	Family member	No. of rare functional variants (not shared)	No. of proband-only variants (revealed)	Gene name	Proven causal gene	Protection quotient	Bandwidth (gigabytes)	Compute (min)	Network (min)
	Trio	Father	185	N/R	N/R	*ACTB*	1 – 2/524 = 99.6%	18.1	1.7	56.7
		Mother	164	N/R	N/R					
		Proband	175	2	*ACTB* *USH2A*					
INTERSECTION (over variants)	Scenario: two hospitals	No. of suspicious variants (not shared)	Total intersecting variants (for patient phenotype comparison follow-up)				Protection quotient	Bandwidth (gigabytes)	Compute (min)	Network (min)
	Washington Baylor	5734 5333	159				1 – 318/11,067 = 97.1%	3.1	0.37	9.4

is the open-ended nature of current analysis. Even the most massive GWASs of any common (complex) disease today explain only a fraction of the observed disease heritability (*3*).

An alternative approach for secure computation is to use fully homomorphic encryption (FHE) (*28*). An FHE scheme enables arbitrary computation on encrypted genomic data (*29*) and also allows one to reuse the same (encrypted) input across multiple computations. However, current implementations of FHE are quite inefficient and do not readily scale for genome-wide computation or for evaluating complicated functions. As a concrete comparison, for the functions we describe here, a conservative estimate shows that our Yao-based solution is at least 5000 to 10,000 times faster compared with the current state-of-the-art FHE scheme (*30*).

Mendelian genomics offers immediate applicability for cryptography. Mendelian diseases affect a sizable fraction of the population and offer a definitive diagnosis for an ever-growing number of diseases. With appropriate modifications, our approach can be extended to complex (multigenic) diseases with a definitive diagnosis. Fewer than 50 multigenic diseases have been diagnosed with certainty to date (*2*). Our approach can search for specific combinations of these genes in affected patients and output a match to any of these. As structured approaches are developed for the discovery of multigenic disease causes, our algorithm can be adapted to support these new approaches.

The computational resources we use to ensure genomic privacy are not negligible, yet modern computers are capable of completing the operation in seconds or minutes. Although no security mechanism is perfectly impenetrable, it is certainly preferable to have one in place (especially if it allows for exact computation). As we learn to provide more definitive genomic diagnoses and as cryptographic tools continue to improve, widespread deployment of computer libraries that implement tools for genomic privacy will encourage more individuals to securely contribute their genomes for the common good.

REFERENCES AND NOTES

1. A. M. Wenger, H. Guturu, J. A. Bernstein, G. Bejerano, *Genet. Med.* **19**, 209–214 (2017).
2. A. M. Gazzo *et al.*, *Nucleic Acids Res.* **44**, D900–D907 (2016).
3. I. M. Nolte *et al.*, *Eur. J. Hum. Genet.* **25**, 877–885 (2017).
4. H. L. Rehm *et al.*, *Genet. Med.* **15**, 733–747 (2013).
5. M. Lek *et al.*, *Nature* **536**, 285–291 (2016).
6. S. B. Ng *et al.*, *Nature* **461**, 272–276 (2009).
7. S. B. Ng *et al.*, *Nat. Genet.* **42**, 30–35 (2010).
8. S. B. Ng *et al.*, *Nat. Genet.* **42**, 790–793 (2010).
9. J. Kaiser, *Science* **354**, 398–399 (2016).
10. M. D. Mailman *et al.*, *Nat. Genet.* **39**, 1181–1186 (2007).
11. L. L. Siu *et al.*, *Nat. Med.* **22**, 464–471 (2016).
12. S. S. Shringarpure, C. D. Bustamante, *Am. J. Hum. Genet.* **97**, 631–646 (2015).
13. A. Regalado, Internet of DNA: A global network of millions of genomes could be medicine's next great advance. *MIT Technol. Rev.* (2015); www.technologyreview.com/s/535016/internet-of-dna/.
14. M. A. Simpson *et al.*, *Nat. Genet.* **43**, 303–305 (2011).
15. J.-B. Rivière *et al.*, *Nat. Genet.* **44**, 440–444 (2012).
16. K. L. McBride *et al.*, *Hum. Mol. Genet.* **17**, 2886–2893 (2008).
17. S. Fenske *et al.*, *Circ. Res.* **109**, 1015–1023 (2011).
18. Y.-F. Huang, B. Gulko, A. Siepel, *Nat. Genet.* **49**, 618–624 (2017).
19. N. Sobreira, F. Schiettecatte, D. Valle, A. Hamosh, *Hum. Mutat.* **36**, 928–930 (2015).
20. A. A. Philippakis *et al.*, *Hum. Mutat.* **36**, 915–921 (2015).

21. Y. Erlich, A. Narayanan, *Nat. Rev. Genet.* **15**, 409–421 (2014).
22. M. S. Riazi, N. K. R. Dantu, L. N. V. Gattu, F. Koushanfar, in *Proceedings of the 2016 IEEE International Symposium on Hardware Oriented Security and Trust (HOST)* (IEEE, 2016), pp. 248–253.
23. Y. Lindell, B. Pinkas, *J. Cryptol.* **22**, 161–188 (2009).
24. C. Hazay, Y. Lindell, *Efficient Secure Two-Party Protocols - Techniques and Constructions* (Information Security and Cryptography Series, Springer, 2010).
25. C. Dwork, in *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming* (Springer, 2006), pp. 1–12.
26. S. Simmons, C. Sahinalp, B. Berger, *Cell Syst.* **3**, 54–61 (2016).
27. F. Chen *et al.*, *Bioinformatics* **33**, 871–878 (2017).
28. C. Gentry, in *Proceedings of the 41st ACM Symposium on Theory of Computing* (ACM, 2009), pp. 169–178.
29. N. Dowlin *et al.*, *Proc. IEEE* **105**, 552–567 (2017).
30. I. Chillotti, N. Gama, M. Georgieva, M. Izabachène, in *Proceedings of the 22nd Annual International Conference on the Theory and Applications of Cryptology and Information* (Springer, 2016), pp. 3–33.

ACKNOWLEDGMENTS

We thank J. Bernstein and members of the Boneh and Bejerano labs for valuable discussions, tools, and project feedback. We also thank Stanford patients and clinicians, as well as the patients and professionals involved in the deposition of the dbGaP (Database of Genotype and Phenotype) sets that we used. This work was supported in part by Stanford Graduate and Computational and Evolutionary Human Genomics Fellowships (K.A.J.), an NSF Graduate Research Fellowship (D.J.W.), a Stanford Interdisciplinary Graduate Fellowship (J.A.B.), Simons and National Science Foundation Fellowships (D.B.), the Stanford Pediatrics Department, Defense Advanced Research Projects Agency, the Packard Foundation, and Microsoft Faculty Fellowships (G.B.). We declare no conflicts of interest. All patient studies were performed in compliance with the Stanford Institutional Review Board. Code is available at https://github.com/dwu4/genome-privacy.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/357/6352/692/suppl/DC1
Materials and Methods
Figs. S1 to S6
References (*31*–*48*)

13 February 2017; accepted 18 July 2017
10.1126/science.aam9710

Deriving genomic diagnoses without revealing patient genomes

Karthik A. Jagadeesh, David J. Wu, Johannes A. Birgmeier, Dan Boneh and Gill Bejerano

Science **357** (6352), 692-695.
DOI: 10.1126/science.aam9710

Sharing data, protecting privacy

Although data-sharing is crucial for making the best use of genetic data in diagnosing disease, many individuals who might donate data are concerned about privacy. Jagadeesh *et al.* describe a solution that combines a protocol from modern cryptography with frequency-based clinical genetics used to diagnose causal disease mutations in patients with monogenic disorders. This framework correctly identified the causal gene in cases involving actual patients, while protecting more than 99% of individual participants' most private variants.

Science, this issue p. 692