Inference Attacks against Kin Genomic Privacy

Erman Ayday | Bilkent University
Mathias Humbert | Swiss Data Science Center

Genomic data poses serious interdependent risks: your data might also leak information about your family members' data. Methods attackers use to infer genomic information, as well as recent proposals for enhancing genomic privacy, are discussed.

Individuals desiring to control their personal data face significant *interdependent privacy risks*—risks that involve the leakage of one's personal data due to data shared by other individuals. With recent advances in whole genome sequencing, genomic data in particular poses serious interdependent privacy risks.

Genomic data has many unique characteristics: it is highly valuable, is an individual's distinctive fingerprint, rarely changes throughout an individual's lifetime, is nonrevocable, and includes sensitive information about an individual (such as disease status or physical characteristics).[1,2] But, the main reason genomic data poses interdependent privacy risks is that it's correlated within family members. Thus, one person's genome-related data (for instance, raw genome, variant call format file, genomic test results, or aggregate statistics) might leak information about the genome-related data of his or her family members.

This issue goes all the way back to the DNA dragnets that first raised serious concerns among privacy advocates. Here, we present recent developments on the information security front, including

- how attackers can infer an individual's genomic data from the partial genomes of his or her family members, background knowledge about genomics (simple statistics, high-order correlations, and so on), and the individual's phenotypic information;
- how attackers can determine an individual's membership in a particular genomic dataset (for example, a beacon) from only the results of basic queries to that dataset and partial genomic knowledge about the individual's family members;
- how attackers can deanonymize the deidentified genomes in a public dataset by using the kinship information; and
- how attackers can efficiently infer kinship from public anonymous genomic databases.

Background

Before discussing these developments in further detail, we introduce the important genomic elements relevant to this article.

Genomic Elements

The vast majority (approximately 99.5 percent) of DNA is similar among human beings. Of the remaining 0.5 percent, the most common variant in the human genome is called a *single nucleotide polymorphism*

1540-7993/17/$33.00 © 2017 IEEE Copublished by the IEEE Computer and Reliability Societies September/October 2017

(SNP). An SNP is a variation of a nucleotide at a specific position in the genome that affects at least 1 percent of individuals in a given population (typically referred to as a *common SNP*). As of November 2016, the Single Nucleotide Polymorphism database (dbSNP; www.ncbi.nlm.nih.gov/projects/SNP) lists approximately 154 million common SNPs in human beings. An SNP, like any other base pair, has two nucleotides. Each nucleotide can take either the major or minor allele. The major allele is the most commonly observed nucleotide in a given population, whereas the minor is the rare nucleotide. If we represent the major allele as B and the minor allele as b, an SNP can take values in $\{BB; Bb; bb\}$, where B and b take values in the alphabet $\{A; T; G; C\}$. SNP values are also known as an individual's genotype.

SNPs are especially sensitive from a privacy perspective because many of these polymorphic positions are associated with severe diseases. For example, carrying particular values at two SNPs (rs7412 and rs429358) on the Apolipoprotein E (ApoE) gene indicates an increased risk for Alzheimer's disease.

Due to genetic inheritance laws, family members share more SNPs than unrelated individuals. Thus, SNPs can be used to infer kinship between two individuals. Moreover, kinship information can infer hidden (or unknown) SNP values of relatives. Also commonly used for kinship inference are *short tandem repeats* (STRs). STRs consist of two to 13 nucleotides repeated numerous times in a row on the DNA strand. For instance, GATAGATAGATA is an STR of period four repeating three times. STRs have a higher mutation rate than other areas of DNA, leading to high genetic diversity.

Reproduction

Mendel's first law of inheritance—the law of segregation—states that alleles are passed independently from parents to child for different meioses (children). Moreover, at each SNP position, the child inherits one allele from the mother and one from the father. Each allele from the parents is randomly selected from their two alleles with probability 0.50. Hence, if the mother has an SNP value of BB and the father has an SNP value of Bb, the child will inherit an SNP equal to BB or Bb, both with probability 0.50. If both parents carry an SNP equal to Bb, then the child's SNP will take a value of BB or bb with probability 0.25, and value Bb with probability 0.50. Finally, given both parents' genomes, the child's genome is independent of all other ancestors' genomes.

One exception to Mendel's law is the Y chromosome. The Y chromosome is inherited (almost) intact along a family's male line. Thus, a father's Y chromosome is the same as his son's Y chromosome. Due to this property, multiple genealogy companies offer services to reunite distant patrilineal relatives by genotyping a few dozen highly polymorphic STRs across the Y chromosome (called Y-STRs).

Another exception to the law of segregation is mitochondrial DNA (mtDNA), which is the DNA located in mitochondria of cells. mtDNA is inherited only from the mother, and hence enables researchers to trace a family's maternal lineage.

Inference Attacks on Kin Genomic Privacy

In this section, we discuss the main threats against kin genomic privacy.

DNA Dragnets

The privacy risks posed by genomic data's collection and use in forensics have been widely discussed in the context of DNA dragnets. DNA dragnets involve collecting tissue or saliva samples from people in a certain region to hunt criminals. The collected biological samples are then used to construct DNA databases. Although collecting such data from suspected criminals or from those who've given their informed consent is acceptable, there are still serious privacy implications.

A main concern about DNA dragnets is the conditions under which law enforcement is legally allowed to collect individuals' biological samples. Under the US Fourth Amendment, law enforcement must have a reasonable suspicion that a person is involved in a crime before requiring a search or seizure. However, the rules for DNA collection are still uncertain. For instance, in Melbourne, Florida, riding a bike at night without two functioning lights could lead to a DNA swab.[3]

Another concern is the duration such samples are kept in DNA databases and whether law enforcement can use the samples for other investigations. In 2015, Maryland's Supreme Court ruled that law enforcement could use DNA voluntarily provided to police investigating one crime to solve another.[3]

Also of concern is using research databases that collect biological samples in criminal investigations—without informing the donors about such use.[4] Such forensic investigations have occurred in Australia, New Zealand, Norway, the UK, and Sweden for criminal identification, disaster victim identification, and paternity identification. A prominent example was the use of Sweden biobank blood samples to investigate the 2003 murder of a Swedish foreign minister.

One last serious privacy concern about DNA dragnets relates to kinship: law enforcement might use an individual's DNA from a DNA database to accuse a family member whose biological sample was never collected. Some US states already allow such familial searching of DNA databases. However, there are

concerns over whether the right to privacy is violated in the process.

DNA technology used by genealogists to identify unknown relatives and DNA dragnets used by law enforcement have been successfully combined to track down criminals. For example, police spent nearly 20 years (starting in the 1970s) chasing the BTK ("bind, torture, and kill") serial killer.[5] Use of DNA in forensics and familial DNA connections finally helped them identify the killer. Although the police already had the suspect's DNA samples from the crime scenes and strong evidence that BTK was a man named Dennis Rader, they didn't have the reasonable doubt necessary to get a DNA swab from Rader. Police learned that Rader's daughter had recently been to the hospital for a pap smear. Thus, via a judge's order (but without the daughter's knowledge), the police received a sample of the daughter's DNA from the hospital, determined the familial match between that sample and the crime scene DNA samples, and eventually caught Dennis Rader.

On one hand, familial search in forensics DNA databases is a powerful tool for the police. Experts state that this technique increases the number of suspects identified through DNA by 40 percent.[3] On the other hand, privacy advocates question the legitimacy of obtaining information with this technique because it turns family members into genetic informants without their knowledge or consent.

Quantifying Kin Genomic Privacy

In previous work, we provided a quantification framework for assessing the effect on kin genomic privacy of family members revealing their genomes.[6] To precisely quantify privacy, we mimicked an adversary who has access to some genome(s) in a given family and wants to infer the genomes of other family members. To do so, the adversary relies on the intergenome correlations (data between relatives); the observed genomic and phenotypic data; and, potentially, intragenome correlations (so-called linkage disequilibrium), typically if a genome is only partially observed. Our efficient inference algorithms were based on belief propagation and graphical models. Belief propagation let us reduce the complexity of computing marginal distributions of random variables from time exponential to linear in the number of considered variables.

Once the belief propagation algorithm output the posterior marginal probabilities given the observed genome(s) and phenotype(s), we quantified the change in genomic privacy with respect to the prior probability distribution given by general population statistics. To do so, we relied on the expected estimation error and success rate (which requires us to know the ground truth, or actual SNP value) and on entropy-based metrics, which measure the adversary's uncertainty and don't require the ground truth.

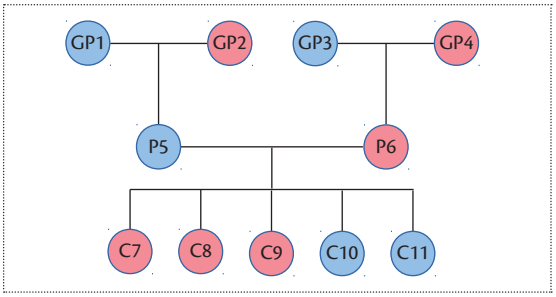

Figure 1. The CEPH/Utah Pedigree 1463 family tree consisting of 11 family members, which includes four grandparents (GP1 to GP4), two parents (P5 and P6), and five children (C7 to C11).[8]

We evaluated the proposed inference attacks and showed their efficiency and accuracy by using real genomic data from CEPH/Utah Pedigree 1463.[7] Specifically, we selected 11 family members: the four grandparents (GP1 to GP4), the two parents (P5 and P6), and the five children (C7 to C11; see Figure 1). We focus here on the results of all common SNPs available on chromosome 1 (approximately 80,000). Table 1 shows the evolution of the expected estimation error and the success rate (the probability of inferring the correct SNP value) given the observation of zero to three different relatives. The three main rows represent the targeted individual (whose genomic data is hidden), and the columns represent the observed genomic data used to infer the hidden, targeted data. Looking at the P5 row, we see that we can decrease the average error by 50 percent by observing only P5's two parents, and by even more if we also observe one of his children. Note that the proportion of SNPs inferred with success greater than 0.90 increases from 20 to 57 percent by observing P5's parents. This proportion increases to 87 percent when seven of his relatives are observed (not shown in table). This clearly demonstrates that genomic privacy can be dramatically damaged by others' sharing behavior.

Effect of High-Order Correlations in the Genome

To analyze the use of high-order correlations in the genome to improve existing work on inference attacks on genomic privacy, we also considered the phenotype–genotype relationships (such as physical traits or disease information).[8] We used the complex correlations in the genome by applying Markov and recombination models between the *haplotypes*—nucleotides on a single chromosome that are so closely linked that they're

Table 1. Absolute and relative levels of genomic privacy of the grandparent (GP1), parent (P5), and child (C7) whose genome is hidden (H), given the observation (∅) of zero to three relatives.

H/O	Error*	∅	P5	P5, GP2	C7, GP2	C7, C8, GP2
GP1	Absolute average error	0.446	0.322	0.309	0.404	0.385
	Relative average error (%)	100	72	69	91	86
	Single nucleotide polymorphisms (SNPs) with success rate >0.90 (%)	20	28	29	23	23
		∅	GP1, GP2	C7, C8	C7, P6	GP1, GP2, C7
P5	Absolute average error	0.480	0.242	0.286	0.312	0.203
	Relative average error (%)	100	50	60	65	42
	SNPs with success rate >0.90 (%)	20	57	38	29	57
		∅	P5	P5, C8	P5, P6	P5, P6, C8
C7	Absolute average error	0.489	0.344	0.301	0.182	0.182
	Relative average error (%)	100	70	62	37	37
	SNPs with success rate >0.90 (%)	20	28	40	64	64

**We use the absolute error to measure the genomic privacy of GP1, P5, and C7 for each individual, the error relative to the initial error (without observing any data) as a percentage, and the proportion of SNPs with a success rate over 0.90. The success rate is the probability of inferring the correct SNP value.*

usually inherited as a unit. Then, similar to existing work,[6] we proposed an efficient graph-based, iterative message–passing algorithm to consider all the aforementioned background information for the inference. Overall, our results show that an attacker's inference power significantly improves by using complex correlations and phenotype information along with information about family bonds.

For evaluation, we focused on 100 neighboring SNPs on the CEPH/Utah Pedigree 1463's DNA sequence on the 22nd chromosome. Using data from the 1000 Genomes Project (www.internationalgenome.org) and HapMap (www.ncbi.nlm.nih.gov/genome/probe/doc/ProjHapmap.shtml), we modeled the genome's higher-order correlations (Markov and recombination models).

Among the 100 SNPs, we randomly hid 50 of the father's SNPs and tried to infer them by gradually increasing the attacker's background information. We also assumed that the attacker knew three of each family member's phenotypes associated with the considered SNPs. We began revealing 50 random SNPs (out of 100) of other family members, starting from the most distant to the father in terms of number of family tree hops. To quantify genomic privacy, we used two metrics: estimation error and entropy.

Figure 2 shows our results for the attacker's error (we achieved similar results for the entropy-based metric). The case of $k = 1$ (Markov chain with order 1 with no phenotype information) represents our previous work.[6] Our results show that high-order correlations and phenotype information contributed significantly to the attacker's inference power. For the Markov chain model, the attacker's inference didn't improve much for orders of Markov chain (k) greater than 3. The recombination model increased the attacker's inference power more than the Markov chain model.

Suppose we're working on a dataset consisting of a trio (father, mother, and child) and trying to infer a particular SNP of the father given the mother's and child's SNPs. Following Mendel's law, if the child is homozygous (carrying two identical nucleotides) in that SNP position, we can easily infer the nucleotide in one strand of the father. However, if both the child and the mother are heterozygous (carrying two different nucleotides) in that SNP position, we can't get any information about the nucleotide passed on from the father to the child.[6]

We can ameliorate this limitation by using haplotype information. Haplotypes are identical by descent (IBD) if they're identical and inherited from a common ancestor. There are several ways to detect IBD.[9,10] Previously, we used Beagle[11] for this and showed IBD's contribution to the inference attack.[12] Beagle allows SNPs to be in linkage disequilibrium (LD) by modeling haplotype frequencies.

By employing this haplotype information, we introduced a new inference attack to find one of the parent's SNPs by using the genomes of the other parent and the children. We used the regions that are inherited together and worked from the idea that if the child's SNPs in a haplotype block aren't coming from the mother's genome, then they're coming from the father's. Then, we deduced that the child's other haplotype is inherited from the father. We evaluated our approach on CEPH/Utah Pedigree 1463 dataset, and showed that accurate inference about the father's SNPs could be accomplished using less data (that is, less genomic data from fewer family members) than previously.[6]

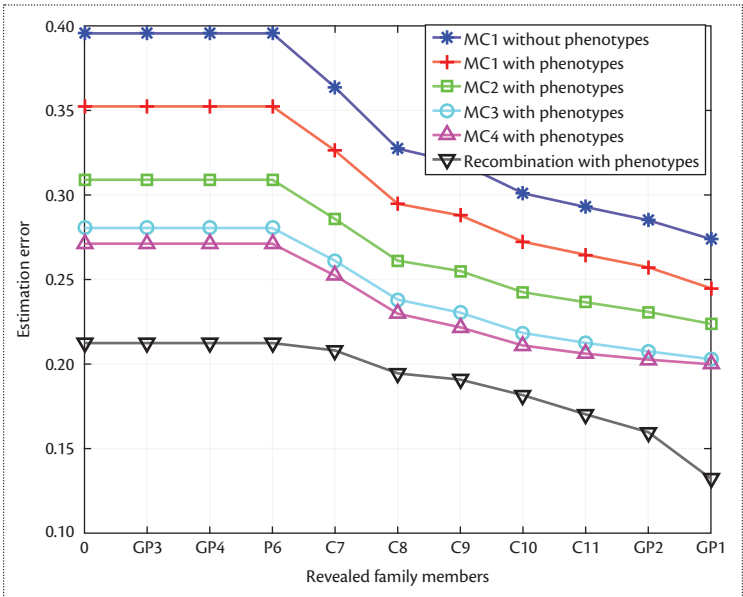

Figure 2. Decrease in father's genomic privacy by attacker's incorrectness. We revealed partial genomes of other family members for different high-order correlation models in the genome. MC is Markov chain model (with different orders).

Membership Inference in Genomic Databases

In 2008, Nils Homer and his colleagues identified an attack against genomic privacy that determined a targeted individual's membership in a genomic database based on summary statistics about this database.[13] By comparing a significant portion of the targeted individual's SNPs with the released statistics, the adversary could infer with high precision whether the individual was a member of the database.

The following year, Sriram Sankararaman and his colleagues proposed another statistical inference method, one based on likelihood ratio, to derive a theoretical bound on the attack's true-positive at a given false-positive rate.[14] They showed that it's possible to detect relatives of the target whose SNPs are available to the adversary. Notably, they found that detecting a target's first-order relative (sibling, child, or parent) requires approximately four times as many SNPs as detecting the target with the same bound on false-positive and false-negative rates. Moreover, they empirically demonstrated that if the adversary has access to approximately 33,000 independent common SNPs, the true-positive rate decreases from 0.95 (for detecting the original individual) to 0.22 for detecting a first-order relative, and 0.03 for a second-order relative, at a false-positive rate of 10^{-3}.

More recently, Suyash Shringarpure and Carlos Bustamante developed an attack against genomic data-sharing beacons.[15] Beacons are webservers that answer allele presence queries such as "Do you have a genome that has a specific nucleotide (A) at a specific genomic position (position 11,272 on chromosome 1)?" with either "yes" or "no." By relying on a likelihood-ratio test, the authors showed that the responses to such queries could be used to reidentify individuals in a beacon.

Moreover, Shringarpure and Bustamante showed that relatives are also prone to such a reidentification attack. Similar to Sankararaman and his colleagues, the authors used a single parameter to model the degree of relatedness (the probability that two individuals share an allele at a single SNP: 1.00 for identical twins, 0.50 for parent–offspring and sibling pairs, 0.25 for first cousins, and so on) and derive the updated likelihood-ratio test as a function of this parameter. Using simulated data, they showed that in a beacon with 1,000 individuals, target reidentification was possible—at a more than 0.95 true-positive rate and 0.05 false-positive rate—with only 5,000 queries; first-order relative reidentification required approximately 40,000 queries. The true-positive rate dropped to 0.50 for second-order relatives, and approximately 0.23 for third-order relatives, with 40,000 queries.

Deanonymizing Publicly Available Genomic Datasets

As discussed, the Y chromosome is (almost) preserved along the male line of a given family. Thus, for communities in which last name is also preserved along the male line, the Y chromosome and last names are correlated. Such correlation can be accessed through public genealogy databases.

Melissa Gymrek and her colleagues recently showed that individuals' last names could be recovered by querying recreational genealogy databases with their Y-STRs.[16] Furthermore, the combination of last name with other auxiliary information such as age and state (which can be easily obtained from public resources) could be used to triangulate the target's identity. Eventually, such triangulation would lead an attacker to link the anonymized genomic data stored on a public repository to the donor's real identity.

The authors used the public genealogy databases Ysearch (www.ysearch.org) and SMGF (www.smgf.org), both of which are free and have built-in search engines. When users input their (or someone else's) Y-STR profile, the database returns the last name of the corresponding donor. Gymrek and her colleagues also assumed that anonymized genomic data (from which they obtained the target's Y-STR profile) is available with the target's birth year and state of residency. Note that the US's Health Insurance Portability and Accountability Act of 1996 (HIPAA) doesn't protect these two pseudoidentifiers.

Finally, the authors determined the target's real identity by entering the target's last name, birth year, and state of residency into online public record search engines. They showed that this combination yielded a median result set (the set containing potential donors of a given anonymized genome) of 12. They also reported five successful surname inferences—in which the anonymized genome's donor could be uniquely identified—from Illumina datasets of three large families that were part of the 1000 Genomes Project, which eventually exposed nearly 50 research participants' identities.

Countermeasures

Here, we briefly discuss some potential countermeasures against these privacy risks.

Cryptography-Based Solutions

Keeping genomic data in encrypted form, instead of making it publicly available, and providing query results only to specific individuals (such as patients, medical centers, or researchers) might mitigate some of the aforementioned attacks. Cryptography-based techniques can protect both kin and personal genomic privacy. To this end, researchers have proposed cryptographic solutions for different query types.

There's been a significant amount of work on privacy-preserving pattern matching and the comparison of genomic sequences. Juan Ramon Troncoso-Pastoriza and his colleagues proposed an algorithm for private string searching on the DNA sequence by using a finite state machine.[17] Their work was revisited by Marina Blanton and Mehrdad Aliasgari, who developed an efficient method for sequence comparison using garbled circuits.[18] Furthermore, Muhammad Naveed and his colleagues proposed a scheme based on functional encryption for privacy-preserving similarity tests on genomic data.[19] Recently, Xiao Shaun Wang and his colleagues proposed an efficient privacy-preserving protocol to find genetically similar patients in a distributed environment.[20]

Other works have focused on private clinical genomics. Emiliano De Cristofaro and his colleagues proposed a secure protocol between two parties that tests genomic sequences without leaking private information about the genomic sequence or the test's nature.[21] Pierre Baldi and his colleagues used private-set intersection to present an effective algorithm for privacy-preserving clinical tests and direct-to-consumer methods on DNA sequences.[22] Rui Wang and his colleagues proposed computing on genomic data by distributing the task between a data provider and consumer through program specialization.[23] Erman Ayday and his colleagues designed a scheme that protects the privacy of users' genomic data while enabling medical units to access the data to conduct medical tests or develop personalized medicine methods.[24] Finally, Zhicong Huang and his colleagues developed an information-theoretical technique to securely store genomic data.[25]

One last line of investigation has explored the use of cryptography-based techniques such as homomorphic encryption, secure hardware, and secure multiparty computation.[26,27]

Differential Privacy–Based Solutions

Cryptography-based techniques help individuals query genomic databases in a privacy-preserving way. However, such solutions don't prevent an attacker from making inferences from the results of such queries. As for cryptographic mechanisms, the techniques for mitigating membership inference were developed to protect personal genomic privacy in general. However, differential privacy, a well-known technique for answering statistical queries in a privacy-preserving manner,[28] can be easily adapted to preserve kin genomic privacy at a lower cost for utility because membership inference is more successful for individuals whose genomic data is known than for their kin.

To prevent such attacks, differential privacy has been used to compose privacy-preserving query mechanisms for genome-wide association study (GWAS) settings.[29,30] Caroline Uhler and her colleagues proposed methods for releasing differentially private minor allele frequencies (MAFs), chi-square statistics, p-values, top-k most relevant SNPs to a specific phenotype, and specific correlations between particular SNP pairs.[29] These methods are notable because traditional

differential privacy techniques would be unsuitable: the number of correlations studied in GWAS is much larger than the number of people in the study. However, differential privacy is typically based on a mechanism that invokes Laplacian noise and, thus, requires a very large number of research participants to guarantee acceptable privacy and utility levels.

Aaron Johnson and Vitaly Shmatikov explained that computing the number of relevant SNPs and the pairs of correlated SNPs is the goal of a typical GWAS.[30] They provided a distance score mechanism to add noise to the output. All relevant queries required by a typical GWAS are supported, including the number of SNPs associated with a disease and the most significant SNPs' locations. Empirical analysis suggests that the new distance score–based, differentially private queries produced better, though still far from acceptable, utility for a typical GWAS. Differential privacy might also be a solution for the beacon attack, with a tradeoff in utility.

Optimization-Based Solutions

Differential privacy techniques perturb the data before releasing it, and cryptographic techniques are generally too inefficient for research settings. To avoid these issues, some individuals might decide to publicly share their data in clear (without encryption), for example, to help medical research progress. In a previous work, we proposed an optimization-based mechanism for reaching a suitable tradeoff between shared SNPs' usefulness and family members' genomic privacy.[31] Optimization-based solutions could potentially mitigate all the attacks we've discussed. The optimization-based solution we discuss subsequently is particularly tailored to inference attacks.

Consider individuals who want to share their genome, yet are concerned about the subsequent privacy risks for themselves and their family. We designed a system that maximizes disclosure utility without exceeding a certain level of privacy loss within a family, considering kin genomic privacy, the family members' personal privacy preferences, the SNPs' privacy sensitivities, the correlations between SNPs, and the SNPs' research utility. Our solution automatically evaluates the privacy risks for all family members and decides which SNPs to disclose. It relies on the quantification framework discussed earlier and combinatorial optimization.

First, we defined a linear optimization problem that aims to maximize the utility of disclosed SNPs. Utility increases linearly with the number of shared SNPs, while satisfying all family members' genomic and health privacy constraints. This problem is very similar to the optimization literature's multidimensional knapsack problem; we relied on the branch-and-bound algorithm to find the optimal SNP subset to be disclosed. Second, we applied a fine-tuning algorithm to account for the impact of intragenome correlations (linkage disequilibrium) on privacy. Our results indicated that, given the current data model, we can protect an entire family's genomic privacy while still making available an appropriate subset of genomic data. The approach's main disadvantage is that the considered optimization problem is nondeterministic polynomial time–complete and doesn't admit any fully polynomial-time approximation scheme. Therefore, we can't consider a significant number of SNPs using this problem.

Future Research Directions

Individuals are increasingly using direct-to-consumer services such as 23andMe, AncestryDNA, and FamilyTreeDNA to obtain their genomic information. Some share this information on public genome-sharing websites such as openSNP.org, mainly to contribute to genomic research. Although most share their genomic data on such platforms in an anonymized way, others either directly reveal their real identities or share sufficient information to cause deanonymization.[16,32] By analyzing the genomic data of such websites' users, attackers might be able to infer family bonds; if at least one family member is identifiable or deanonymized, attackers might be able to reconstruct the actual family tree along with their genomic data.

Although this poses a serious privacy risk for contributors to anonymized genomic datasets, these datasets are crucial to genomic research. To find the balance between privacy and utility, an optimization-based solution, similar to the one we discussed, could be used. By selectively hiding dataset participants' SNPs, such an optimization-based technique would also hide familial relationships between the donated genomes and maximize the utility of the data shared by the donors.

Other types of biomedical data are becoming increasingly available, such as epigenomic or transcriptomic data. In particular, DNA methylation, one of the most important epigenomic elements, was recently shown to be reidentifiable through genotype inference,[33] because parts of the DNA methylation are influenced by the genome. These correlations between DNA methylation and the genome imply the existence of interdependent privacy risks for relatives' DNA methylation data. Therefore, it's crucial to precisely quantify these interdependent risks and analyze whether they appear beyond the parts of the DNA methylation that are correlated with the genome.

The kinship-related privacy implications of genomic data will only continue to grow as genomics gain importance and more people get their DNA sequenced.

Thus, it's crucial that we consider and implement appropriate protective mechanisms when using individuals' genomic data in various applications. ■

Acknowledgments

Erman Ayday was supported by funding from the European Union Horizon 2020 Research and Innovation Programme (Marie Skodowska-Curie grant 707135) and the Scientific and Technological Research Council of Turkey, TUBITAK (grant 115C130).

References

1. M. Naveed et al., "Privacy in the Genomic Era," *ACM Computing Surveys*, vol. 48, no. 1, 2015; doi.org/10.1145/2767007.
2. Y. Erlich and A. Narayanan, "Routes for Breaching and Protecting Genetic Privacy," *Nature Rev.*, vol. 15, no. 6, 2014, pp. 409–421.
3. L. Kirchner, "DNA Dragnet: In Some Cities, Police Go from Stop-and-Frisk to Stop-and-Spit," *ProPublica*, 12 Sept. 2016; www.propublica.org/article/dna-dragnet-in-some-cities-police-go-from-stop-and-frisk-to-stop-and-spit.
4. V. Dranseika, J. Piasecki, and M. Waligora, "Forensic Uses of Research Biobanks: Should Donors Be Informed?," *Medicine, Health Care, and Philosophy*, vol. 19, no. 1, 2016, pp. 141–146.
5. E. Nakashima, "From DNA of Family, a Tool to Make Arrests," *Washington Post*, 21 Apr. 2008; www.washingtonpost.com/wp-dyn/content/article/2008/04/20/AR2008042002388.html.
6. M. Humbert et al., "Quantifying Interdependent Risks in Genomic Privacy," *ACM Trans. Privacy and Security*, vol. 20, no. 1, 2017, pp. 1–31.
7. R. Drmanac et al., "Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays," *Science*, vol. 327, no. 5961, 2010, pp. 78–81.
8. I. Daznabi et al., "An Inference Attack on Genomic Data Using Kinship, Complex Correlations, and Phenotype Information," to be published in *IEEE/ACM Trans. Computational Biology and Bioinformatics*, 2017.
9. B.L. Browning and S.R. Browning, "A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals," *Am. J. Human Genetics*, vol. 84, no. 2, 2009, pp. 210–223.
10. J.M. Rodriguez, S. Batzoglou, and S. Bercovici, "An Accurate Method for Inferring Relatedness in Large Datasets of Unphased Genotypes via an Embedded Likelihood-Ratio Test," *Proc. 17th Int'l Conf. Research in Computational Molecular Biology* (RECOMB 13), 2013, pp. 212–229.
11. B.L. Browning and S.R. Browning, "A Fast, Powerful Method for Detecting Identity by Descent," *Am. J. Human Genetics*, vol. 88, no. 2, 2011, pp. 173–182.
12. F. Balci et al., "A New Inference Attack against Kin Genomic Privacy," *Proc. Privacy-Aware Computational Genomics* (PRIVAGEN 15), 2015; www.cs.bilkent.edu.tr/~erman/pubs/PrivaGen_inference.pdf.
13. N. Homer et al., "Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays," *PLoS Genetics*, vol. 4, no. 8, 2008; doi.org/10.1371/journal.pgen.1000167.
14. S. Sankararaman et al., "Genomic Privacy and Limits of Individual Detection in a Pool," *Nature Genetics*, vol. 41, no. 9, 2009, pp. 965–967.
15. S.S. Shringarpure and C.D. Bustamante, "Privacy Risks from Genomic Data-Sharing Beacons," *Am. J. Human Genetics*, vol. 97, no. 5, 2015, pp. 631–646.
16. M. Gymrek et al., "Identifying Personal Genomes by Surname Inference," *Science*, vol. 339, no. 6117, 2013; doi.org/10.1126/science.1229566.
17. J.R. Troncoso-Pastoriza, S. Katzenbeisser, and M. Celik, "Privacy Preserving Error Resilient DNA Searching through Oblivious Automata," *Proc. 14th ACM Conf. Computer and Communications Security* (CCS 07), 2007, pp. 519–528.
18. M. Blanton and M. Aliasgari, "Secure Outsourcing of DNA Searching via Finite Automata," *Proc. 24th Ann. IFIP WG 11.3 Working Conf. Data and Applications Security and Privacy* (DBSec 10), 2010, pp. 49–64.
19. M. Naveed et al., "Controlled Functional Encryption," *Proc. ACM SIGSAC Conf. Computer and Communications Security* (CCS 14), 2014, pp. 1280–1291.
20. X.S. Wang et al., "Efficient Genome-Wide, Privacy-Preserving Similar Patient Query Based on Private Edit Distance," *Proc. ACM SIGSAC Conf. Computer and Communications Security* (CCS 15), 2015, pp. 492–503.
21. E. De Cristofaro et al., "Secure Genomic Testing with Size- and Position-Hiding Private Substring Matching," *Proc. 12th ACM Workshop Privacy in the Electronic Society* (WPES 13), 2013, pp. 107–118.
22. P. Baldi et al., "Countering GATTACA: Efficient and Secure Testing of Fully-Sequenced Human Genomes," *Proc. ACM SIGSAC Conf. Computer and Communications Security* (CCS 11), 2011, pp. 691–702.
23. R. Wang et al., "Privacy-Preserving Genomic Computation through Program Specialization," *Proc. ACM Conf. Computer and Communications Security* (CCS 09), 2009, pp. 338–347.
24. E. Ayday et al., "Protecting and Evaluating Genomic Privacy in Medical Tests and Personalized Medicine," *Proc. 12th ACM Workshop Privacy in the Electronic Society* (WPES 13), 2013, pp. 95–106.
25. Z. Huang et al., "Genoguard: Protecting Genomic Data against Brute-Force Attacks," *Proc. IEEE Symp. Security and Privacy* (SP 15), 2015; doi.org/10.1109/SP.2015.34.

26. M. Kantarcioglu et al., "A Cryptographic Approach to Securely Share and Query Genomic Sequences," *IEEE Trans. Information Technology in Biomedicine*, vol. 12, no. 5, 2008, pp. 606–617.
27. M. Canim, M. Kantarcioglu, and B. Malin, "Secure Management of Biomedical Data with Cryptographic Hardware," *IEEE Trans. Information Technology in Biomedicine*, vol. 16, no. 1, 2012, pp. 166–175.
28. C. Dwork, "Differential Privacy," *Proc. 33rd Int'l Conf. Automata, Languages and Programming* (ICALP 06), 2006, pp. 1–12.
29. C. Uhler, A. Slavkovic, and S.E. Fienberg, "Privacy-Preserving Data Sharing for Genome-Wide Association Studies," *J. Privacy and Confidentiality*, vol. 5, no. 1, 2013, pp. 137–166.
30. A. Johnson and V. Shmatikov, "Privacy-Preserving Data Exploration in Genome-Wide Association Studies," *Proc. 19th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining* (KDD 13), 2013, pp. 1079–1087.
31. M. Humbert et al., "Reconciling Utility with Privacy in Genomics," *Proc. 13th Workshop Privacy in the Electronic Society* (WEPS 14), 2014, pp. 11–20.
32. M. Humbert et al., "De-anonymizing Genomic Databases Using Phenotypic Traits," *Proc. 15th Privacy Enhancing Technologies Symp.* (PETS 15), 2015, pp. 99–114.
33. M. Backes et al., "Identifying Personal DNA Methylation Profiles by Genotype Inference," *Proc. 38th IEEE Symp. Security and Privacy* (SP 17), 2017; doi.org/10.1109/SP.2017.21.

Erman Ayday is an assistant professor of computer science at Bilkent University. His research interests include privacy-enhancing technologies (including big data and genomic privacy), data security, and trust and reputation management. Ayday received a PhD in electrical and computer engineering from Georgia Tech. He's a member of IEEE and ACM. Contact him at erman@cs.bilkent.edu.tr.

Mathias Humbert is a senior data scientist at the Swiss Data Science Center, ETH Zurich, and École Polytechnique Fédérale de Lausanne (EPFL). His research interests include genomic privacy, privacy in online social networks, and location privacy. Humbert received a PhD in interdependent privacy from EPFL. He's a member of IEEE and ACM. Contact him at mathias.humbert@epfl.ch.

myCS Read your subscriptions through the myCS publications portal at **http://mycs.computer.org**

ReliabilitySociety

Executive Committee (ExCom) Members: Jeffrey Voas, President; Dennis Hoffman, Sr. Past President, Christian Hansen, Jr. Past President; Pierre Dersin, VP Technical Activities; Pradeep Lall, VP Publications; Carole Graas, VP Meetings and Conferences; Joe Childs, VP Membership; Alfred Stevens, Secretary; Bob Loomis, Treasurer

Administrative Committee (AdCom) Members:
Joseph A. Childs, Pierre Dersin, Lance Fiondella, Carole Graas, Samuel J. Keene, W. Eric Wong, Scott Abrams, Evelyn H. Hirt, Charles H. Recchia, Jason W. Rupe, Alfred M. Stevens, Jeffrey Voas, Marsha Abramo, Loretta Arellano, Lon Chase, Pradeep Lall, Zhaojun (Steven) Li, Shiuhpyng Shieh

http://rs.ieee.org

The IEEE Reliability Society (RS) is a technical society within the IEEE, which is the world's leading professional association for the advancement of technology. The RS is engaged in the engineering disciplines of hardware, software, and human factors. Its focus on the broad aspects of reliability allows the RS to be seen as the IEEE Specialty Engineering organization. The IEEE Reliability Society is concerned with attaining and sustaining these design attributes throughout the total **life cycle. The Reliability Society has the management, resources, and administrative and technical structures to develop and to provide technical information via publications, training, conferences, and technical library (IEEE Xplore) data to its members and the Specialty Engineering community. The IEEE Reliability Society has 28 chapters and members in 60 countries worldwide.**

The Reliability Society is the IEEE professional society for Reliability Engineering, along with other Specialty Engineering disciplines. These disciplines are design engineering fields that apply scientific knowledge so that their specific attributes are designed into the system / product / device / process to assure that it will perform its intended function for the required duration within a given environment, including the ability to test and support it throughout its total life cycle. This is accomplished concurrently with other design disciplines by contributing to the planning and selection of the system architecture, design implementation, materials, processes, and components; followed by verifying the selections made by thorough analysis and test and then sustainment.

Visit the IEEE Reliability Society website as it is the gateway to the many resources that the RS makes available to its members and others interested in the broad aspects of Reliability and Specialty Engineering.

IEEE