Mathias Humbert*, Kévin Huguenin, Joachim Hugonot, Erman Ayday, and Jean-Pierre Hubaux

De-anonymizing Genomic Databases Using Phenotypic Traits

Abstract: People increasingly have their genomes sequenced and some of them share their genomic data online. They do so for various purposes, including to find relatives and to help advance genomic research. An individual's genome carries very sensitive, private information such as its owner's susceptibility to diseases, which could be used for discrimination. Therefore, genomic databases are often anonymized. However, an individual's genotype is also linked to visible phenotypic traits, such as eye or hair color, which can be used to re-identify users in anonymized public genomic databases, thus raising severe privacy issues. For instance, an adversary can identify a target's genome using known her phenotypic traits and subsequently infer her susceptibility to Alzheimer's disease. In this paper, we quantify, based on various phenotypic traits, the extent of this threat in several scenarios by implementing de-anonymization attacks on a genomic database of OpenSNP users sequenced by 23andMe. Our experimental results show that the proportion of correct matches reaches 23% with a supervised approach in a database of 50 participants. Our approach outperforms the baseline by a factor of four, in terms of the proportion of correct matches, in most scenarios. We also evaluate the adversary's ability to predict individuals' predisposition to Alzheimer's disease, and we observe that the inference error can be halved compared to the baseline. We also analyze the effect of the number of known phenotypic traits on the success rate of the attack. As progress is made in genomic research, especially for genotype-phenotype associations, the threat presented in this paper will become more serious.

Keywords: Privacy; Genomics; De-anonymization; Graph matching

DOI 10.1515/popets-2015-0020
Received 2015-02-15; revised 2015-04-23; accepted 2015-05-15.

1 Introduction

Due to the decreasing cost of genome sequencing, more and more people have their genotypes sequenced. The most popular direct-to-consumer service provider, 23andMe,[1] has already genotyped more than 800,000 individuals' DNA. This new availability of genomic data is paving the way for revolutionary medical progress. Among other benefits, access to genomic data enables personalized medicine (e.g., personal drug dosing) and early diagnosis of severe genetic diseases (such as Alzheimer's or Parkinson's). In order to help research progress (typically genome-wide association studies) and benefit from personalized treatment, a significant number of genotyped individuals share their genomic data online (on platforms like OpenSNP[2] [13] or Personal Genome Project[3]) or on semi-public databases (e.g., at hospitals or national research institutions).

However, genomic data carries very sensitive information about its owner, such as a predisposition to certain diseases, future physical conditions, and kinship, which could lead to discrimination or familial tragedies if it is not properly and securely handled [3, 37]. The genomic-privacy issue is exacerbated by the fact that genomic data is non-revocable and highly correlated between close relatives [18]. As a consequence, whether publicly or semi-publicly disclosed, genomic data is often shared without identifying information (e.g., name). However, it has been shown that quasi-identifying attributes (such as ZIP code, or birth date) can be used in order to re-identify participants, notably in the Personal Genome Project [34]. In this case, the background knowledge (or auxiliary information) needed for the attack was obtained through voter lists. More recently, researchers have de-anonymized Y-chromosome short tan-

***Corresponding Author: Mathias Humbert:** EPFL, Lausanne, Switzerland, E-mail: mathias.humbert@epfl.ch
Kévin Huguenin: LAAS-CNRS, Toulouse, France, E-mail: kevin.huguenin@laas.fr
Joachim Hugonot: EPFL, Lausanne, Switzerland, E-mail: joachim.hugonot@epfl.ch
Erman Ayday: Bilkent University, Ankara, Turkey, E-mail: erman@cs.bilkent.edu.tr
Jean-Pierre Hubaux: EPFL, Lausanne, Switzerland, E-mail: jean-pierre.hubaux@epfl.ch

[1] https://www.23andme.com/
[2] https://opensnp.org
[3] http://www.personalgenomes.org/

dem repeats (STRs) by using (as auxiliary information) notably Y-chromosome genealogical websites that provide the surname of male individuals corresponding to the genomic data [14]. This attack relies upon the comparison of Y-chromosome STRs that are currently not included in the genotypes provided by most direct-to-consumer genetic testing providers (such as 23andMe).

In this paper, we propose new de-anonymization attacks[4] that makes use of only the most common piece of genomic information that is output today by major direct-to-consumer providers: the single nucleotide polymorphims (SNPs). The attack relies upon the fact that our SNPs are intrinsically linked to our phenotypic traits (such as eye color, blood type, or genetic diseases) and that genomic research progress provides us with more information about these links. For instance, the relationship between SNPs and phenotypes is increasingly used in forensics for reconstructing facial composites from DNA information [7, 8].[5] Therefore, if an adversary has access to phenotypic traits (e.g., visible traits) of an identified individual, he can use known correlations between phenotypic traits and genomic data to identify the genotype of this individual in a genomic database and to infer other sensitive information (such as predispositions to severe diseases) by using the de-anonymized genomic data, as illustrated in Figure 1. The adversary also have access to anonymized genotypes through a collaborative genome-sharing platform (such as the Personal Genome Project) and want to de-anonymize them by relying upon phenotypic information gathered on online social networks (OSNs). The matching between OSNs and genomic profiles is (even) easier if, for example, the ZIP code is available with the genomic profiles, thus enabling the OSN profiles to be filtered before the matching attack. The adversary might also want to match different online identities, e.g., OpenSNP profiles that contain genomic data with PatientsLikeMe[6] profiles that contain phenotypic information (mainly health condition, such as diseases).

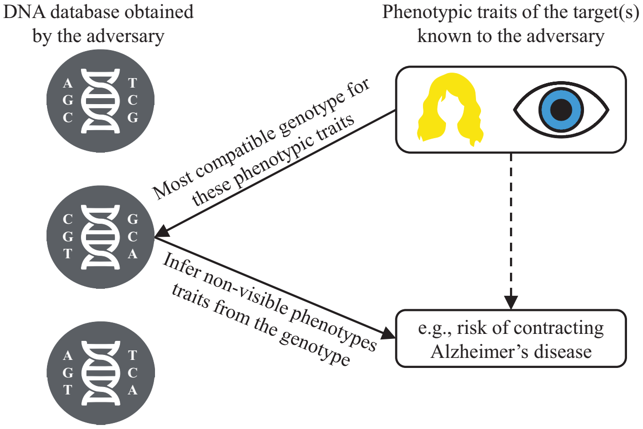

DNA database obtained by the adversary
Phenotypic traits of the target(s) known to the adversary
A G C
T C G
C G T
G C A
A G T
T C A
Most compatible genotype for these phenotypic traits
Infer non-visible phenotypes traits from the genotype
e.g., risk of contracting Alzheimer's disease

Fig. 1. Illustration of the identification attack: The adversary identifies the genotypes of a target individual from some of her visible phenotypic traits and uses the de-anonymized genotype to infer her susceptibility to Alzheimer's disease.

More specifically, we study two de-anonymization attacks: (i) the identification attack, where the adversary wants to identify the genotype (among multiple genotypes) that corresponds to a given phenotype, and (ii) the perfect matching attack, where the adversary wants to match multiple phenotypes to their corresponding genotypes. We rely upon analytical tools for maximizing the matching likelihood in both attacks, and we assume two types of background knowledge: one that makes use of existing genetic knowledge from the association between SNPs and phenotypic traits (unsupervised approach), and another that learns the genomic-phenotypic statistical relationships from datasets containing both data types (i.e., genomic and phenotypic). Our experimental results show, with a database of 80 participants, in the identification attack, a proportion of 13% correct matches in the supervised case, and 5% in the unsupervised case. These results constitute a significant improvement: they outperform the considered baseline by a factor of eight and three, respectively. When the database size decreases to 10, the attack success increases to around 44% in the unsupervised case, and 52% in the supervised case. We also evaluate the adversary's ability to predict the predisposition to Alzheimer's disease of the database participants. With 10 participants, the average error of the adversary is halved when using the identification attack.

In the perfect matching attack, the proportion of correct matches is slightly better than in the identification attack: 16% in the supervised case and 8% in the unsupervised case. For a database of size 10, this proportion increases to 65% and 58%, respectively. With this size, the proportion of correct match is around

[4] These belong to *identity tracing attacks* in the categorization proposed by Erlich and Narayanan [10]. The novelty of the attack presented in this paper lies in the observed data (i.e., phenotypic traits) and the considered probabilistic relationships (i.e., genotype-phenotype), as we describe below.

[5] Such techniques are used in practice; for instance the police department of Columbia, SC, USA, recently released DNA-based facial composites generated with the Parabon Snapshot™ DNA Phenotyping Service [31].

[6] https://www.patientslikeme.com/

four times higher than the baseline for both supervised and unsupervised approaches. We also evaluate the impact of the distinguishability between two individuals on the success of the perfect matching attack. Our results clearly show that the more distinguishable two individuals are, the more likely their genomic (or phenotypic) data will be de-anonymized. This leads us to conclude that the threat on genomic privacy posed by our de-anonymization attacks will become even more serious in the near future, when more SNP-trait association information is discovered by genomic researchers, and available to the adversary. Finally, we propose various countermeasures for mitigating the performance of de-anonymization attacks against genomic databases.

The rest of the paper is organized as follows. In Section 2, we introduce the important concepts used in the paper. In Section 3, we describe the system and adversarial model, including various examples of de-anonymization attacks. In Section 4, we present the analytical model and techniques that we will use for de-anonymizing genomic data. In Section 5, we thoroughly evaluate the performance of the proposed attacks by relying upon real data. In Section 6, we suggest different mechanisms for reducing the genomic-privacy risks. In Section 7, we summarize the related work, before concluding in Section 8.

2 Background

In this section, we introduce relevant notions about genotypes, phenotypes and their relationship, which will be helpful for the understanding of the rest of the paper.

The human genome is encoded in double stranded DNA molecules consisting of two complementary polymer chains. Each chain consists of simple units called nucleotides. Each nucleotide is assigned a letter from the set {A,C,G,T}, and a human genome consists of approximately three billion pairs of letters.

Approximately 99.5% of any two individuals' genomes are exactly the same, and the remaining 0.5% is referred to as the genetic variation. The most common genetic variant (position in the genome that holds a nucleotide that varies between individuals) in the human population is the single nucleotide polymorphism (SNP).

In general, there are two different alleles (nucleotides) observed at a given SNP position: (i) the major allele is the most frequently observed nucleotide, and (ii) the minor allele is the rare nucleotide. Thus, each SNP is assigned a minor and a major allele frequency, among which the minor allele frequency is the smaller of the two. Furthermore, each SNP position includes two alleles (i.e., two nucleotides) and everyone inherits one allele of every SNP position from each of her parents. If an individual receives the same allele from both parents, her SNP is said to be *homozygous*. Her SNP is *homozygous major* if it inherits two major alleles, and *homozygous minor* if it inherits two minor alleles. If a SNP inherits a different allele from each parent's SNP (one minor and one major), it is called *heterozygous*.

Today, there are approximately 50 million approved (by the research community) SNPs in the human population [29] and each individual carries on average 4 million variants (i.e., SNPs carrying at least one minor allele) out of this 50 million. DNA encodes the proteins synthesized by an organism. As these proteins affect our physical appearance (e.g., Melanosomal proteins influence the color of an individual's skin), an individual's DNA is linked to her phenotypic traits. Therefore, SNPs have a direct influence on our physical attributes (e.g., hair color, eye color, blood type) but also on our predispositions to various diseases. The fact that two monozygotic twins look almost alike provides clear evidence of the influence of DNA on our physical (notably visible) attributes. The relationship between genotype and phenotype is generally determined by association studies over a population that is split into case and control groups. Each SNP contributes to the disease susceptibility (or physical attribute) in a different amount and the contribution amount of each SNP is determined through these association studies. Furthermore, some of the SNPs contribute to the development of a disease, whereas some are protective. Note that environmental factors can have more influence than SNPs on the development of certain phenotypic traits, especially those that are related to diseases.

In the rest of the paper, we call *genotype*, respectively *phenotype*, the set (or vector) of *SNPs*, respectively of *(phenotypic) traits*. Also, we use the adjective *genomic*, respectively *phenotypic*, to mention anything related to the genotype and the phenotype, respectively.

3 System and Adversarial Model

The adversary is assumed to have access to two distinct datasets: (i) a set $\mathcal{G} = \{\mathbf{g}_1, \mathbf{g}_2, ..., \mathbf{g}_n\}$ of genotypes of n different individuals, where $\mathbf{g}_i = (g_{i,1}, \dots, g_{i,s})$ is a vector containing the SNP values of individual i, with

$g_{i,j} = \{0, 1, 2\}$[7] and (ii) a set $\mathcal{P} = \{\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_m\}$ of phenotypes of m individuals, where $\mathbf{p}_i = (p_{i,1}, \ldots, p_{i,t})$ is a vector containing the values of phenotypic traits of i. Note that $p_{i,j} \in \mathcal{P}_j$, where $\mathcal{P}_j$ is the set of values trait j can take. For instance, if trait j represents eye color, we could have $\mathcal{P}_j = \{$"brown","blue","green"$\}$. The genomic dataset can be gathered online, on platforms such as 1000 Genomes Project, the Personal Genome Project, OpenSNP, or be leaked to the adversary through a major security breach, e.g., of a medical database. The phenotypic traits can be collected online, via online social networks (OSNs) (e.g., Facebook, PatientsLikeMe), or by having access to medical databases, or even by learning about them in real life. Not all participants have all their s SNP values known, or all t traits accessible. We consider the general problem where some auxiliary information is provided possibly with the phenotypic data (e.g., name), and some other auxiliary information goes potentially with the genomic data (e.g., ZIP code). These auxiliary sets are assumed to be disjoint, thus are not used for matching both datasets. Therefore, in order to learn more information about the participants in these separate databases (e.g., de-anonymize the genomic data), the adversary seeks to link the genomic dataset with the phenotypic dataset by relying only upon the genomic-phenotypic statistical relationships.

This system and adversarial model enable us to represent various attack scenarios. For instance, this model can represent a curious entity (e.g., an insurer) who wants to identify the genotype of a targeted individual in a database, knowing some of his phenotypic traits (e.g., gathered on an OSN), in which case $m = 1$. It can also model a curious entity that has access to a dataset of anonymized genotypes online indicating some risks to certain severe diseases and who wants to know to whom they belong in order to discriminate them (access to insurance, or price of premiums). It can also represent the situation of a curious person (e.g., an hospital IT staff) who has access to a database with patients' diseases and another database with patients' anonymized genomes, and who wants to de-anonymize these genomes. It can also be that the adversary has access to an identified genotype on OpenSNP and wants to use it to identify the corresponding user on PatientsLikeMe (where potentially all medical conditions, their evolution and treatments are available). Note that the adversary can use side information such as the individuals' ZIP code or the age to narrow down the set of possible individuals in the genomic database; this would result into, typically, a few hundreds of individuals at most. Another possible attack (not considered in this paper), which relies on the same line of reasoning as the aforementioned attacks, is to directly infer a individual's genotype based on some of her phenotypic traits and on the genotype-phenotype relationships.

In this work, we consider two types of background knowledge available to the attacker for linking the genomic and phenotypic datasets. In the first model, we assume the adversary to have access to a genomic knowledge database that stores information about the relations between genomic and phenotypic data (e.g., SNPedia[8]). Such databases typically depict qualitative relationships between SNPs and phenotypic traits. In the second model, we consider a stronger adversary who has access to population statistics about the genomic-phenotypic relations. This represents the scenario where the adversary can construct his knowledge from an existing dataset that contains both genomic and phenotypic data about participants. The former model will be referred to as the *unsupervised* approach, whereas the latter model will be referred to as the *supervised* approach. Note finally that population allele frequencies of SNPs are publicly accessible, thus are also part of the background knowledge.

4 De-anonymization Attacks

We present here in detail the formalization of two de-anonymization attacks. First, assuming the adversary knows the phenotypic traits of a targeted individual, it wants to identify the target's genotype among n other genotypes in a dataset. We refer to this as the *identification attack.* Second, assuming the adversary has access to one genomic and one phenotypic dataset containing the same n individuals, it wants to match genotypes to their corresponding phenotypes. We refer to this as the *perfect matching attack.*

In the identification attack, the adversary needs to rank the n genotypes it has access to by decreasing value of the phenotype likelihood given each possible genotype. If $k \leq t$ traits of target x are observed, and if the

[7] the value 0 represents a homozygous major SNP (e.g., AA), 1 a heterozygous SNP (AT), and 2 a homozygous minor SNP (TT).

[8] http://www.snpedia.com/

number of SNPs j available in the n genotypes vary between 1 and s, the adversary chooses the genotype $\mathbf{g}_i$ that maximizes the following likelihood:

$$\mathbf{P}(\mathbf{p}_x \,|\, \mathbf{g}_i) = \mathbf{P}(p_{x,1}, \dots, p_{x,k} \,|\, g_{i,1}, \dots, g_{i,j})\,, \tag{1}$$

where $1 \leq j \leq s$. Under reasonable assumptions, this likelihood can be simplified as follows:

$$\begin{aligned}\mathbf{P}(p_{x,1}, \dots, p_{x,k} \,|\, g_{i,1}, \dots, g_{i,j}) &\simeq \prod_{l=1}^{k} \mathbf{P}(p_{x,l} \,|\, g_{i,1}, \dots, g_{i,j}) \\ &= \prod_{l=1}^{k} \mathbf{P}(p_{x,l} \,|\, \{g_{i,r}\}_{r \in \mathcal{R}_l}) \\ &\simeq \prod_{l=1}^{k} \prod_{r \in \mathcal{R}_l} \mathbf{P}(p_{x,l} \,|\, g_{i,r}), \end{aligned} \tag{2}$$

where $\mathcal{R}_l$ is the set of SNPs that are relevant to trait l. The first equality in (2) is an approximation that follows from the conditional independence between phenotypic traits given the genotype.[9] The second equality comes from the independence between the traits and the SNPs that do not affect these traits; and the last equality is an approximation. This approximation is made essentially because the relationships between traits and genotypes are provided on SNPedia in single SNP-single trait combinations, like $\mathbf{P}(p_{x,l} \,|\, g_{i,r})$. We keep the same conditional probabilities in the supervised case for consistency and efficiency in the learning phase. We intend to learn the exact conditional probabilities $\mathbf{P}(p_{x,l} \,|\, \{g_{i,r}\}_{r \in \mathcal{R}_l})$ in a supervised approach in future work.

Note that, in case a relevant SNP $g_{i,r}$ is missing in some of the n genotypes, the probability $\mathbf{P}(p_{x,l} \,|\, g_{i,r})$ then becomes $\mathbf{P}(p_{x,l})$. The prior probability of the phenotypic trait can be computed by relying upon the law of total probability:

$$\mathbf{P}(p_{x,l}) = \sum_{\forall g_{i,r} \in \{0,1,2\}} \mathbf{P}(p_{x,l} \,|\, g_{i,r})\, \mathbf{P}(g_{i,r}), \tag{3}$$

where $\mathbf{P}(g_{i,r})$ is provided by public population statistics. Note that in (3), for presentation simplicity, we considered the case where only one SNP is relevant to the phenotypic trait. The formula generalizes straightforwardly to more relevant SNPs.

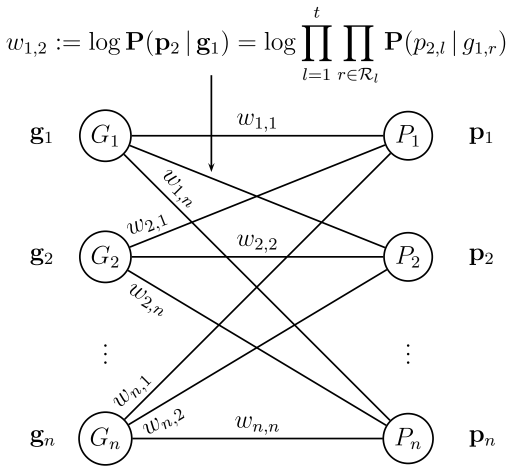


Fig. 2. Complete weighted bipartite graph with n vertices on the lefthand side representing the individuals' genotypes and n other vertices on the righthand side representing their phenotypes. The weight $w_{i,j}$ is the log-likelihood of phenotype $\mathbf{p}_j$ given genotype $\mathbf{g}_i$.

In the perfect matching attack, the goal of the adversary is to assign precisely one genotype to one phenotype, such that the resulting n assignments maximize the product of the likelihoods $\prod_{i=1}^{n} \mathbf{P}\big(\mathbf{p}_{\sigma(i)} \,|\, \mathbf{g}_i\big)$ over all $n!$ assignments σ between size-n sets $\mathcal{G}$ and $\mathcal{P}$. Hence, the assignment σ^* that maximizes likelihood is

$$\sigma^* = \arg\max_{\sigma} \prod_{i=1}^{n} \prod_{l=1}^{t} \prod_{r \in \mathcal{R}_l} \mathbf{P}\big(p_{\sigma(i),l} \,|\, g_{i,r}\big). \tag{4}$$

Simply put, this problem is finding a perfect matching on a weighted bipartite graph, with n vertices on one side representing the n different genotypes, and n vertices on the other side representing the n phenotypes, as shown in Figure 2. A weight is assigned to every edge of the complete bipartite graph. We define the weight $w_{i,j}$ between a genotype vertex G_i and a phenotype vertex P_j as the log-likelihood between between genotype $\mathbf{g}_i$ and phenotype $\mathbf{p}_j$:

$$w_{i,j} := \log \mathbf{P}(\mathbf{p}_j \,|\, \mathbf{g}_i) = \log \prod_{l=1}^{t} \prod_{r \in \mathcal{R}_l} \mathbf{P}(p_{j,l} \,|\, g_{i,r}), \tag{5}$$

and we solve the following optimization problem:

$$\begin{aligned}\sigma^* &= \arg\max_{\sigma} \sum_{i=1}^{n} w_{i,\sigma(i)} && (6) \\ &= \arg\max_{\sigma} \sum_{i=1}^{n} \log \prod_{l=1}^{t} \prod_{r \in \mathcal{R}_l} \mathbf{P}\big(p_{\sigma(i),l} \,|\, g_{i,r}\big) && (7) \\ &= \arg\max_{\sigma} \log \prod_{i=1}^{n} \prod_{l=1}^{t} \prod_{r \in \mathcal{R}_l} \mathbf{P}\big(p_{\sigma(i),l} \,|\, g_{i,r}\big) && (8)\end{aligned}$$

The formulation in (8) enables us to maximize the sum of the weights instead of their product. Many existing algorithms can find the solution to this optimization

[9] Note that we make the assumption that the environment does not create dependencies between traits.

problem in polynomial time. Here we use the blossom algorithm that finds the maximum weight assignment in $O(n^3)$ [11]. We choose this algorithm because it has the smallest complexity and it can be applied to general graphs. The construction of the bipartite graph, i.e., computation of its weights, takes $O(tn^2)$, given that the number of relevant SNPs per phenotypic trait is constant. Finally, as the logarithmic function is monotically increasing, the optimal assignment σ^* derived in (8) must be the same as the one in (4). Note that maximum weight assignment algorithms have also been used in different contexts, such as the de-anonymization of location traces [32] and of users of anonymous communications [35].

Note that, in the case the set of phenotypic traits or SNPs disclosed by individuals is not the same, we simply assign a constant c to the conditional probability $\mathbf{P}(p_{j,l} \mid g_{i,r})$ of those whose SNP r or trait l is unknown. This will not change the assignment as this constant will be present in the weights of the n edges connected to the genotype vertex G_i (if SNP r is missing) or to the phenotype vertex P_j (if trait l is missing).

We quantify the success of our attacks with different metrics. The proportion of pairs correctly matched reflects the correctness of the de-anonymization attacks in general. In the identification attack target at individual j, the proportion of correct matches is equal to 1 if

$$j = \arg\max_i \mathbf{P}(\mathbf{p}_j \mid \mathbf{g}_i) \tag{9}$$

and 0 otherwise. In the perfect matching attack, we measure the proportion of correct matches, i.e. the ratio between the number of pairs correctly matched (i.e., where $\sigma^*(i) = i$), and the total number of matched pairs n.

In the identification attack, we also evaluate the error of the adversary that tries to infer the susceptibility to a certain disease d. To do so, we compute the average distance between the actual value of the SNPs of the individual j and the SNP values of the individual i that most likely matches the phenotype. We sum over all SNPs that contribute to disease d (whose indices are in set $\mathcal{D}$), and normalize by the number of such SNPs contributing to d and by the maximum L_1 distance between two SNPs (which is equal to 2):

$$\frac{1}{2\,|\mathcal{D}|} \sum_{k \in \mathcal{D}} \|g_{i,k} - g_{j,k}\|_1 \tag{10}$$

Note that if the target's phenotype is correctly mapped to the corresponding genotype, the error is null.

5 Evaluation

In this section, we report on our data-driven evaluation of the de-anonymization attacks presented in the previous section. First, we describe our dataset (its collection, processing and statistics), then we present and analyze the results of our evaluation.

5.1 Dataset Collection and Processing

In order to evaluate the success of the de-anonymization attack presented in the previous section, we collected a dataset of genomic-phenotypic data from OpenSNP in late 2014. OpenSNP is an online platform where users can upload their raw genotype data obtained from direct-to-consumer genetic testing companies such as 23andMe and FTDNA. Together with their genotypes, users can share the values of some of their phenotypic traits. Although the different traits are specified by OpenSNP (e.g., eye color), the value of the traits are manually specified by the users in free-text. We downloaded a raw dump of OpenSNP genomic-phenotypic data in October 2014. We filtered out bogus data (e.g., duplicated, corrupted, empty genotype or phenotype) and, for the sake of homogeneity, we focused on the genomic data obtained from 23andMe. This left us with the data of 818 individual users out of 1137 entries in the raw dump. Note that for some users, the values of some SNPs and phenotypic traits are not known; this is because some SNPs are not tested by 23andMe and because users specify only a subset of the phenotypic traits that appear on OpenSNP.

To obtain statistical association data between SNPs and phenotypic traits, which is used to assign the genotype-phenotype compatibility scores $\{w_{i,j}\}$ to the edges of the bipartite graph, we followed two different approaches: (1) an unsupervised approach in which only qualitative association data, such as (`rs7495174`: AG, "blue eyes more likely"), is available, typically from knowledge databases populated by experts, and (2) a supervised approach in which the SNP-trait associations are *learned* from existing annotated datasets in the form of aggregated statistics such as the proportion of individuals with brown eyes and a value CC for SNP `rs8028689`. In the unsupervised approach, described in Section 5.2.1, we relied on the qualitative association data from SNPedia, based on which we define probabilistic association models for the relevant SNP-trait pairs (according to SNPedia). In the supervised approach, described in Section 5.2.2, we computed the association statistics on a subset of the annotated OpenSNP data, but only for the SNP-trait pairs noted as relevant on SNPedia. We also collected from SNPedia general information about the distribution of the SNP values, for the considered SNPs, across the world population. This information is used as prior distribution in the identification attack when a SNP value is not specified (see Formula 3).

Based on the list of SNPs tested by 23andMe, the list of traits specified on OpenSNP (and the proportion of users who specified a value for these traits) and the association data available on SNPedia, we narrowed down the sets of SNPs and traits considered in our evaluation. In the unsupervised case, we obtained a list of 8 phenotypic traits (deriving from 6 high-level OpenSNP traits, e.g., the "group O", "group AB or B" and "rhesus" traits all deriving from the blood type) with 21 associated SNPs and the form of the sexual chromosomes (i.e., XX or XY). In the supervised case, as we do not need qualitative association information (which are learned from the data—we only use relevance association information), we considered another 4 traits and 14 additional SNPs. The complete list of the traits and SNPs we considered is given in Table 1.

As on OpenSNP the values of the phenotypic traits are provided by the users in free-text, the phenotypic data had consistency issues (e.g., due to the use of different capitalization schemes) and subjectivity issues. For instance, the variety of eye colors was quite high in our dataset, with more than fifty distinct values, including very precise information such as "green with amber burst and gray outer ring". In addition, some of the phenotypic values used on OpenSNP differed from those used on SNPedia. Therefore, we manually translated each possible phenotypic value from OpenSNP to one of the values used on SNPedia, thus narrowing down the set of possible values and reducing the subjectivity of the provided information. Figure 3 shows, in the form of pie charts, the distribution of some of the phenotypic values across our final dataset.

To build our final dataset, we further screened the users based on the following criteria: At least 75% of the phenotypic traits and SNPs considered in the unsupervised case are specified,[11] the SNPs associated with the susceptibility to Alzheimer's disease are specified (i.e., `rs7412` and `rs429358`), no two users have the same genomic or phenotypic information across all the considered SNPs and traits. The reason behind the last criterion is that two individuals with the exact same genomic or phenotypic values *cannot* be distinguished from each other based on the considered SNPs and traits. Note that when the number of considered SNPs and traits increases, so does the uniqueness of the individuals.[12] There exist more than one set of individuals that satisfy these criteria, as an individual can be replaced with another individual with the same genomic information but different phenotypic information (and vice versa). Therefore, we considered such 20 distinct subsets (of 80 individuals)–totaling 94 unique individuals; in our evaluation, we present the results aggregated over the different subsets.

5.2 Experimental Settings and Results

We evaluate the different de-anonymization attacks described in Section 4 in various scenarios, both with an unsupervised and a supervised approach for learning the SNP-trait associations. To do so, we implemented the aforementioned attacks in Python; for the graph maximum-weight matching algorithm, we made use of the implementation from the NetworkX[13] library. We quantify the success of the attack with respect to the two following metrics: (1) the proportion of correctly identified individuals and (2) the accuracy of the susceptibility score (wrt. Alzheimer's disease) computed from the genotypes matched to the individual's phenotype (in the identification attack). Note that for the susceptibility metric, we focus on Alzheimer's disease and SNPs `rs7412` and `rs429358` that significantly affect an individual's chance of having Alzheimer's by the age of 80 [18].

5.2.1 Unsupervised Case

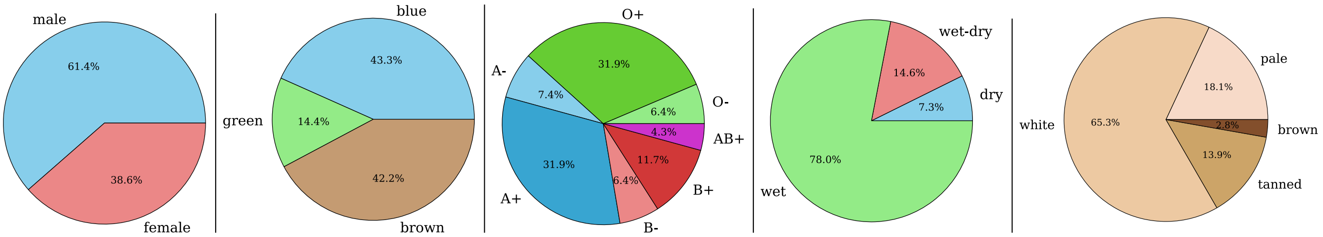
Based on the qualitative association information obtained from SNPedia, we define a probabilistic model

[10] according to Fitzpatrick's scale (see https://en.wikipedia.org/wiki/Fitzpatrick_scale).

[11] by specified, we mean specified with a valid value.

[12] To build our dataset, we removed only 14 individuals for uniqueness reasons (most of the removed individuals were so because they specified too few SNPs/phenotypes). Keeping such duplicates in our dataset would slightly degrade the performance of the attacks. Note that, as the number of SNPs/phenotypes increases, profiles will become unique.

[13] https://networkx.github.io/



(a) Sex **(b)** Eye color **(c)** Blood type **(d)** Earwax type **(e)** Skin color*

Fig. 3. Distribution of some of the phenotypic values for the considered traits in our final dataset. Traits marked with a '*' are considered only in the supervised case.

for the SNP-trait compatibility scores (representing the conditional probabilities of traits given SNPs). Table 2 (see page 16) shows sample probabilistic association models. We first distinguish between the associations that are (quasi-)deterministic, specifically those related to the sex and the blood types and those that are non-deterministic (i.e., for which the genotype influences the phenotype but not with certainty). For the deterministic associations, we introduce a parameter α (close to 1; $\alpha = 1 - 10^{-6}$ in our evaluation) and we set the compatibility score to α for the correct SNP-trait association and $1 - \alpha$ for the incorrect ones.[14] For instance $\mathbf{P}(\text{“male”} \mid \text{XY}) = \alpha$ and $\mathbf{P}(\text{“male”} \mid \text{XX}) = 1 - \alpha$ (see Table 2a, page 16 for the example of blood type). If no association information about a specific SNP value-trait value pair is provided, we use a uniform distribution. For example, we can see in Table 2a that, if SNP rs505922 has value CT, it does not affect the blood type according to the association information available on SNPedia. For the non-deterministic associations, we sort the k possible values and we use a geometric scale (that becomes linear by using the log-likelihood) to translate the qualitative scale into a quantitative one: We set the compatibility score to β_k for the most compatible value of the trait, β_k^2 for the second most compatible value and β_k^k for the less compatible value, where β_k is the solution of the equation $x + x^2 + \cdots + x^k = 1$ as the scores must sum to 1 (see Table 2b, page 16).

5.2.2 Supervised Case

In the supervised case, we build the SNP-trait association models based on the data (we estimate the conditional probabilities on the entire dataset, i.e., the 94 unique individuals, as the dataset is too small to perform a proper cross-validation). For each SNP-trait pair listed in Table 1, we set the compatibility scores to the proportions observed in our dataset. For instance, in the case “O” vs. “Not O” blood type trait, we compute the proportion of individuals with SNP value CC for the SNP rs505922 who have a blood type “O”. By doing so, we build matrices of compatibility scores with the same format as in the unsupervised case. The rest of the identification/perfect matching attack is the same in the unsupervised and supervised cases.

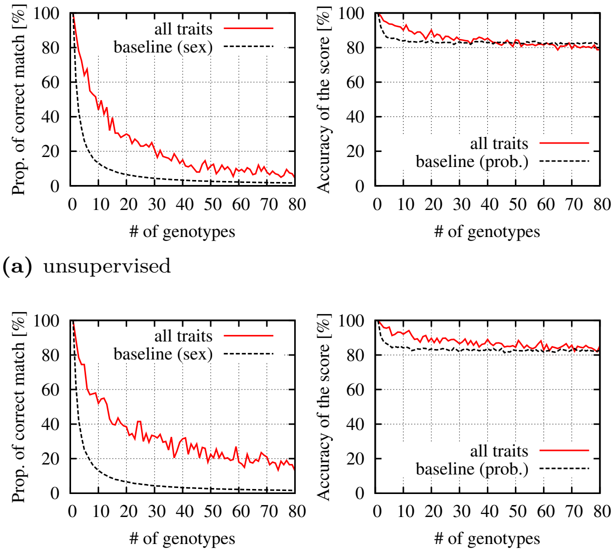
5.2.3 Results

We first look at the success of an identification attack that targets a single phenotype in a genomic database composed of multiple genotypes (at most 80, i.e., the size of our subsets of distinct individuals). This scenario corresponds to the case where an adversary knows that the genotype of a specific individual (whose traits are partially known to the adversary) is present in a given public database; and the adversary tries to identify the genotype of the targeted individual. In this experiment, we assume that all the phenotypic traits of the targeted individual are known to the adversary—the effect of the number of known traits is analyzed in a different experiment described below. We evaluate the success of the adversary for different values of the size of the genomic database, and we show the results, aggregated over all the individuals from the 20 subsets, in Figure 4. It can be observed that, unsurprisingly, the supervised approach achieves better results than the unsupervised approach, and also that the more genotypes there are, the higher the correct-match ratio between the supervised and unsupervised approach is. Indeed, when n increases, the proportion of correct matches in the super-

[14] We need all the probabilities to be non-zero for algorithmic reasons.

vised approach is more than twice as high (i.e., twice as bad, in terms of privacy) as the one in the unsupervised approach. For instance, for a database of 80 genomes, the proportion of correct matches is around 13% in the supervised case vs. 5% in the unsupervised case. In order to gain insight on the performance of the attack on larger databases, we ran experiments with n=210 (this was made possible by lowering the admission threshold to 55% of specified SNPs/phenotypes, to the detriment of the quality of the data); we obtained a proportion of 5.2% of correct matches for the identification attack in the supervised case. Overall, the performance of the identification attack is relatively high, in absolute values and compared to a simple baseline that picks uniformly at random among the individuals of the same sex: The proportion of correct matches of this baseline is 1.6%. We chose this baseline as it is the straightforward genomic-oblivious approach: Once the adversary has ruled out the genotypes that do not match the target's sex, it has no further information to discriminate the remaining genotypes. For a database size of 10 (which corresponds, for instance, to a scenario where an adversary tries to identify the genotype of an individual among DNA samples collected in a room), the proportion of correct matches is around 44% in the unsupervised case (52% in the supervised case), whereas the baseline only achieves 13% of correct matches. Note that the results are relatively high in the unsupervised case, despite the fact that it relies solely on rough association data provided by the SNPedia knowledge base. The result of this approach outperforms by more than three times the baseline and the performance of the supervised approach is four times better than the baseline. Our non-optimized implementation of the perfect-matching attack took 2.7 seconds (on average) to complete on a high-end laptop for $n = 80$.

For the susceptibility score of Alzheimer's disease, the baseline score consists in the expected distance between the actual values of SNPs `rs7412` and `rs429358` of the target and the probabilistic distribution of these two SNPs computed over the same gender population as the target. For a database of size 10, the inaccuracy of the susceptibility score is divided by two between the baseline and our identification attack (dropping from 16% to 8% in the unsupervised case, and from 15% to 7%). This demonstrates that the identification attack can be successfully used to infer private information from an individual's (identified) genotype, especially with small genomic databases for which the proportions of correct matches are relatively high.



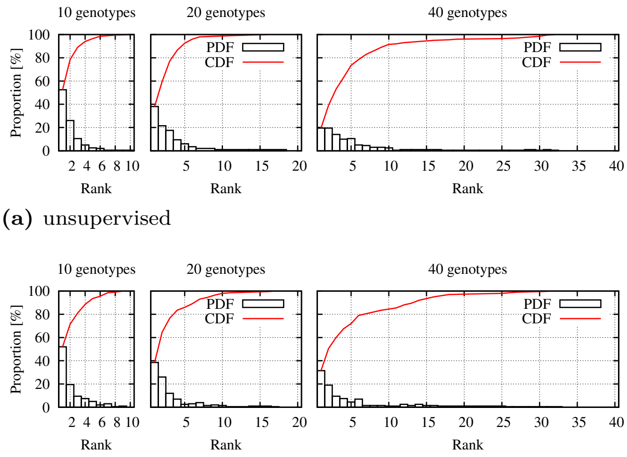
(a) unsupervised

(b) supervised

Fig. 4. Performance of the identification attack for a single individual in a genomic database in the (a) unsupervised and (b) supervised cases, with respect to the proportion of correct matches (left) and the accuracy of the susceptibility score for Alzheimer's disease (right). The size of the database varies from 1 to 80.

In the case of an identification attack that targets a single phenotype, the genotypes in the database can be sorted by decreasing order of likelihood. The matched genotype is the first in the sorted list. To gain insight into the performance of the identification attack, we look at the rank of the target individual's genotype in the sorted list, for different values of the size of the genomic database. The experimental probability density function and the cumulative distribution function of the rank are depicted in Figure 5, for different sizes of the genomic database (i.e., {10,20,40}). It can be observed that in the cases where the target individual's genotype is not in first position, which corresponds to an incorrect match, it often appears at the very beginning of the sorted list of genotypes. For example, in the supervised case with a genomic database of size 20, the target individual's genotype appears in the first 2 elements of the list in 65% of the cases and in the first 5 (out of 20) in 86% of the cases. Again, the results are better in the supervised case, especially when the number of genotypes increase.

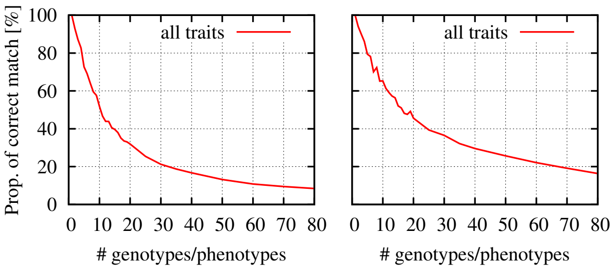
We now consider the perfect matching attack in which the adversary tries to match n genotypes to n phenotypes. We first look at the success of the attack in terms of the number of phenotypes that are matched to the correct genotypes for different sizes of the genomic-phenotypic database (as in Figure 4), and we show the



(a) unsupervised

(b) supervised

Fig. 5. Rank of the target individual's genotype in the list of genotypes of size {10,20,40}, sorted by decreasing order of compatibility in the (a) unsupervised and (b) supervised cases.



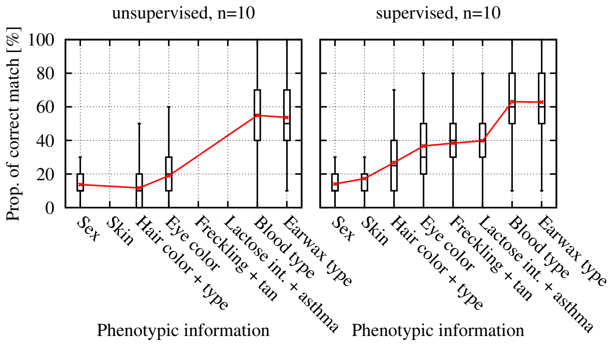
(a) unsupervised **(b)** supervised

Fig. 6. Performance of the perfect-matching attack for the genotypes/phenotypes of 2 to 80 individuals in the (a) unsupervised and (b) supervised cases, with respect to the proportion of correct matches.

results in Figure 6. It can be observed that the performance is very similar to the one of the identification attack, with an even slightly higher proportion of correct matches. For instance, for databases of 80 genomes/phenotypes, the proportion of correct matches is around 16% in the supervised case vs. 8% in the unsupervised case. For a database size of 10, the proportion of correct matches is around 58% in the unsupervised case (65% in the supervised case). We see here, too, that the supervised approach outperforms the unsupervised attack, and we note that the more participants there are in the databases, the higher the success ratio (which tends to 2 when n increases) between the two approaches is. This means that the more participants there are in both databases, the more relevant and helpful is the supervised approach.

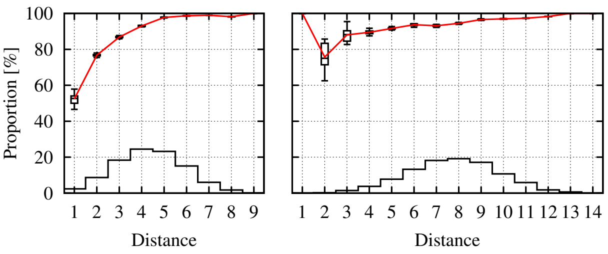
We now look at the effect of the level of knowledge of the adversary about the target individual (number of phenotypic traits known) on the performance of the perfect matching attack. To do so, we sort the different phenotypic traits by increasing "levels of intimacy". By intimacy, we mean the closeness between the adversary and the targeted individuals needed for the adversary to know a specific trait of the target. For instance, the sex of an individual is often common knowledge, while the earwax type is very intimate information. The eye color and the hair type and color also require a low level of intimacy as they can be observed on a picture, posted on a online social network for instance. We sort the different traits by increasing levels of intimacy as shown in Table 1. We acknowledge the fact that this ordering is somewhat arbitrary and arguable; however, it is needed to represent the knowledge of the adversary as a one-dimensional variable. We plot the results, in the form of boxplots showing the first, second (i.e., median) and third quartiles as well as the confidence intervals and the average (red solid line), in Figure 7, for a genomic-phenotypic database of 10 individuals. The labels on the x-axis are cumulative: the label "Skin" means that the adversary knows the sex *and* the skin color of the target individuals. Note that in the unsupervised case, some of the traits are not used. It can be observed that the performance increases with the adversary's level of knowledge. We also notice that (quasi-)deterministic traits (e.g., blood type) provide substantial improvements in the proportion of correct matches. In the supervised case, the adversary achieves a performance of almost 40% success with only visible phenotypic traits that can be observed on a picture. The small decreases in performance in the unsupervised case are caused by the noise in our dataset and the limited reliable association information available on SNPedia. Finally, we note that in both unsupervised and supervised approaches, the proportion of correct matches with all phenotypes is around four times higher than the proportion with the sex information only (which constitutes a baseline).

Finally, we analyze the distinguishability between individuals and its effect on the performance of the attack. To do so, we introduce the notion of dissimilarity between any two individuals. We express this dissimilarity as the minimum value between the Hamming distance on the phenotypes of the individuals and the Hamming distance on their genotypes. As there are fewer phenotypic traits than SNPs, this metric most often represents the dissimilarity between two phenotypes. Figure 8 shows how the proportion of correct



(a) unsupervised (b) supervised

Fig. 7. Performance of the perfect-matching attack for the genotypes/phenotypes of 10 individuals vs. the level of phenotypic knowledge of the adversary about the target individual in the (a) unsupervised and (b) supervised cases, with respect to the proportion of correct matches.



(a) unsupervised (b) supervised

Fig. 8. Performance of the perfect-matching attack with the genotypes/phenotypes of two individuals vs. the distance (or dissimilarity) between these two individuals in the (a) unsupervised and (b) supervised case, with respect to the proportion of correct matches. The stairstep lines represent the proportion of pairs of individuals (in our dataset) with a given distance. The supervised approach contains higher distance between two individuals because it considers more traits than the unsupervised one.

matches evolves with the dissimilarity (distance) between two individuals. We clearly notice that the less similar two individuals are, the more likely they are to be de-anonymized correctly. This means that the genomic-privacy situation will be worsened if the adversary considers more traits and SNPs. Note that, when distance is 1, we get 100% correct matches in the supervised case because of the small number of samples in our dataset (only a few pair samples have a distance of 1, compared to thousands for higher distances).

Our results demonstrate the serious de-anonymization threat currently posed to individuals sharing their SNPs in genomic databases. Our identification and perfect matching attacks outperform the genomic-oblivious baselines by three to four times. Our evaluation also brings up an important point that is that the more distinguishable (genetically or phenotypically) two individuals are, the more likely they are to be de-anonymized. This clearly means that the more SNP-trait associations there are, the more successful will the de-anonymization attacks will be. And it is very likely that, with the progress of genomic research, knowledge databases like SNPedia will grow in size and precision, thus enabling successful attacks against larger databases.

5.3 Limitations

The purpose of our evaluation is to assess the potential of the attacks described in the paper. Although the results provided in the previous section are sufficient to demonstrate that the threat is real, our evaluation and our datasets still have some limitations that we discuss below. A first limitation is the limited size of our dataset (although the size of the dataset used in our evaluation corresponds to practical attack scenarios), which is caused by the low quality of the OpenSNP dataset in terms of the proportion of specified phenotypic traits. As part of future work, we intend to collect larger datasets and perform further experiments on them. Another limitation is the fact that, in our supervised approach, we estimated the conditional probabilities on the entire dataset. The main reason for this is that splitting the dataset into training and testing sets would have further reduced the number n of genotypes/phenotypes considered in the evaluation of the attack. As part of future work, we will study the effect of the training set size on the performance of the attack by taking a rigorous cross-validation approach in the supervised case. Moreover, in order to assess the extent to which the results of the supervised approach can be generalized to other datasets, we intend to collect multiple datasets and run the supervised attack, trained on a given dataset, on a different one.

6 Countermeasures

There are various techniques that can be relied upon for reducing the risk of de-anonymization through genotype-phenotype matching. First of all, upstream protection of genomic data by cryptographic means would dramatically thwart any attempt to de-anonymize this data [4]. It has been proposed to use homomorphic encryption and private set intersection for providing personalized medical tests while preserving privacy of genomic data [5, 6]. Although encryption of genomic data does not reduce much utility and efficiency in healthcare, such cryptographic techniques probably add too much overhead in genomic research [21].

A simpler method for preventing de-anonymization attacks is to split the genomic data into several subsets, making sure that there is no linkage disequilibrium (i.e., probabilistic dependencies between different SNPs of the same genome) between these subsets, so that the records of the different datasets cannot easily be matched. In this way the adversary could not link the various parts of genomic data with each other, thus dramatically reducing the performance of the de-anonymization attacks. Nevertheless, any association study between phenotypes and genotypes could still be carried out, but on smaller parts of the genome.

Another simple technique for thwarting de-anonymization attacks is to selectively share the SNPs; in particular hide those related to phenotypic traits. By removing the SNPs associated with visible traits, we could already significantly decrease the risk of de-anonymization attack that relies on pictures, e.g., gathered on online social networks.

Although the use of noise on aggregated statistical results in the context of genome-wide association studies (for ensuring differential privacy) has not provided very good utility [20, 36], the addition of noise directly on genomic data could be more successful. Following the idea of geo-indistinguishability proposed in the context of location privacy [2], we could add some noise to the genomic data such that the mechanism provides enough indistinguishability between different genomic sequences. As illustrated in Figure 8, the more distinguishable two genotypes are, the more likely they are to be de-anonymized. However, we should control the amount of noise added such that it does not reduce much the accuracy of statistical outcomes and/or personal utility, depending on the use case. We intend to study and evaluate in detail some of the aforementioned privacy-preserving mechanisms in future work.

7 Related Work

Anonymization was one of the first mechanisms proposed to protect the genomic privacy of participants in genetic databases. Unfortunately, the removal of quasi-identifying attributes (e.g., birth date or ZIP code) has been proven ineffective for protecting the anonymity of participants in such databases [12, 15, 16].

For instance, genomic variants on the Y-chromosome being correlated with last names (for males), these can be inferred using public genealogy databases. With further effort (e.g., using voter registration forms) the complete identity of the individual can also be revealed [14]. Also, unique features in patient-location visit patterns in a distributed healthcare environment can be used in publicly available records to link the genomic data to the identity of the individuals [25]. Furthermore, it is shown that Personal Genome Project (PGP) participants can be identified based on their demographics, without using any genomic information [34].

The identity of a participant of a genomic study can also be revealed by using a second sample, that is, part of the DNA information from the individual and the results of the corresponding clinical study [9, 12, 17, 19, 39]. For this reason, even a small set of SNPs of the individual might be sufficient as the second sample. For example, it is shown that as few as 100 SNPs are enough to uniquely distinguish one individual from another [23]. Homer *et al.* [17] prove that the presence of an individual in a case group can be determined by using aggregate allele frequencies and his DNA profile. Homer's attack demonstrates that it is possible to identify a participant in a GWAS study by analyzing the allele frequencies of a large number of SNPs. Wang et al. [39] show a higher risk where individuals can actually be identified from a relatively small set of statistics such as those routinely published in GWAS papers. In particular, they show that the presence of an individual in the case group can be determined based upon the pairwise correlation (i.e., linkage disequilibrium) among as few as a couple of hundred SNPs. Whereas the methodology introduced in [17] requires on the order of 10,000 SNPs (of the target individual), this new attack requires only on the order of hundreds of SNPs.

In another recent study [12], Gitschier shows that a combination of information, from genealogical registries and a haplotype analysis of the Y-chromosome collected for the HapMap Project, enables the prediction of the last names of a number of individuals from the HapMap database. Thus, releasing (aggregate) genomic data is currently banned by many institutions due to this privacy risk. In [40], Zhou *et al.* study the privacy risks of releasing aggregate genomic data. They propose a risk-scale system for classifying aggregate data and a guide for the release of such data.

Several papers have studied phenotype prediction from genomic data, notably as a means of tracing identity. A thorough review of methods for predicting phenotypes from genomic data is provided in [22]. Two studies show that age prediction is feasible from DNA information derived from blood samples [30, 41]. Several genome-wide studies report the influence of genomic data on height [1], body mass index (BMI) [26], eye color [38], and facial shape [8, 24]. Although variabilities of phenotypic traits is currently explained to a small extent by genomic differences, the aforementioned papers clearly demonstrate that genetic knowledge about the relationship between phenotype and genomic data is quickly expanding. Therefore, we can expect that the matching attack evaluated in this paper would become successful with many more individuals in the future.

De-anonymization attacks have been used to jeopardize the anonymity of participants in other contexts. A de-anonymization attack against a large Netflix database successfully re-identified Netflix records of known users by relying upon IMDB as background knowledge [27]. Narayanan and Shmatikov also propose de-anonymizing users of an online social network (Twitter in their case) by making use of another online social network (Flickr) as the source of auxiliary information [28]. Another attack shows that location traces could be de-anonymized by relying upon the social graphs of the traces' owners [33]. In the context of anonymous communication, perfect matching attacks have been proven effective to de-anonymize mixing rounds [35].

Our work differs from previous work as it studies de-anonymization attacks against genomic or phenotypic datasets by leveraging the associations between the genomic and phenotypic data. Moreover, it relies only upon the most common variants currently provided by the major direct-to-consumer genetic testing providers, thus upon the variants that are most accessible online. Finally, it provides a thorough evaluation of the factors that play a role in the success of the de-anonymization attacks, and it paves the way for more investigations on the risky relationship between genomic data and phenotypes.

8 Conclusion and Future Work

In this work, we have thoroughly evaluated two new de-anonymization attacks, by making use of two types of background knowledge. We observe that the supervised approach outperforms the unsupervised approach, and that the success ratio between the two approaches increases with the number of participants, tending to two with 80 participants. We also notice that the proportion of correct matches of our identification attack is three to eight times higher than the baseline. We have demonstrated a decrease of 50% of inference error on the predisposition to Alzheimer's disease with a database of size 10. We notice a slight and unexpected increase of correct matches in the perfect matching attack compared to the identification attack. In particular, the proportion of perfect correct matches reaches 16% with the supervised approach. Our results clearly show the extent of the threat of de-anonymization attacks that rely upon genomic-phenotypic statistical relationships. Finally, our results demonstrate that the more distinguishable two individuals are, the more successful the perfect matching is. This leads us to conclude that the matching risk will continuously increase with the progress of genomic knowledge, which raises serious questions about the genomic privacy of participants in genomic datasets. We should also recall that, once an individual's genomic data is identified, the genomic privacy of all his close family members is also potentially threatened.

In future work, we intend to enhance the matching algorithm and performance of the de-anonymization attacks by learning the conditional probabilities of traits given all their relevant SNPs together as well as the joint probabilities for correlated SNPs, in the supervised approach. We also plan to implement the privacy-preserving mechanisms proposed as countermeasures, and study their effectiveness. Finally, we will evaluate the performance of the attacks on larger datasets.

9 Acknowledgments

Parts of this work were carried out while Kévin Huguenin and Erman Ayday were with EPFL.

References

[1] H. L. Allen, K. Estrada, G. Lettre, S. I. Berndt, M. N. Weedon, F. Rivadeneira, C. J. Willer, A. U. Jackson, S. Vedantam, S. Raychaudhuri, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317):832–838, 2010.

[2] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi. Geo-indistinguishability: Differential privacy

for location-based systems. In *CCS'13: Proc. of the 2013 ACM Conf. on Computer and Communications Security*, pages 901–914, 2013.

[3] E. Ayday, E. De Cristofaro, J. Hubaux, and G. Tsudik. The chills and thrills of whole genome sequencing. *IEEE Computer Magazine*, 2015.

[4] E. Ayday, J. L. Raisaro, U. Hengartner, A. Molyneaux, and J.-P. Hubaux. Privacy-preserving processing of raw genomic data. In *DPM'13: Proc. of the 8th Int'l Workshop on Data Privacy Management*, pages 133–147, 2013.

[5] E. Ayday, J. L. Raisaro, J.-P. Hubaux, and J. Rougemont. Protecting and evaluating genomic privacy in medical tests and personalized medicine. In *WPES'13: Proc. of the 12th ACM Workshop on Privacy in the Electronic Society*, pages 95–106, 2013.

[6] P. Baldi, R. Baronio, E. De Cristofaro, P. Gasti, and G. Tsudik. Countering GATTACA: Efficient and secure testing of fully-sequenced human genomes. In *CCS'11: Proc. of the 18th ACM Conf. on Computer and Communications Security*, pages 691–702, 2011.

[7] P. Claes, H. Hill, and M. D. Shriver. Toward DNA-based facial composites: Preliminary results and validation. *Forensic Science International: Genetics*, 13:208–216, 2014.

[8] P. Claes, D. K. Liberton, K. Daniels, K. M. Rosana, E. E. Quillen, L. N. Pearson, B. McEvoy, M. Bauchet, A. A. Zaidi, W. Yao, et al. Modeling 3D facial shape from DNA. *PLoS Genetics*, 10(3):e1004224, 2014.

[9] D. Clayton. On inferring presence of an individual in a mixture: a bayesian approach. *Biostatistics*, 11(4):661–673, 2010.

[10] Y. Erlich and A. Narayanan. Routes for breaching and protecting genetic privacy. *Nature Reviews Genetics*, 15(6):409–421, 2014.

[11] Z. Galil. Efficient algorithms for finding maximum matching in graphs. *ACM Computing Surveys (CSUR)*, 18(1):23–38, 1986.

[12] J. Gitschier. Inferential genotyping of Y chromosomes in Latter-Day Saints founders and comparison to Utah samples in the HapMap project. *American Journal of Human Genetics*, 84(2):251–258, 2009.

[13] B. Greshake, P. E. Bayer, H. Rausch, and J. Reda. openSNP–A Crowdsourced Web Resource for Personal Genomics. *PLoS ONE*, 9(3):e89204, Mar. 2014.

[14] M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich. Identifying personal genomes by surname inference. *Science*, 339(6117):321–324, 2013.

[15] M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich. Identifying personal genomes by surname inference. *Science*, 339(6117):321–324, 2013.

[16] E. C. Hayden. Privacy protections: The genome hacker. *Nature*, 497:172–174, 05 2013.

[17] N. Homer, S. Szelinger, M. Redman, D. Duggan, and W. Tembe. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genetics*, 4, Aug. 2008.

[18] M. Humbert, E. Ayday, J.-P. Hubaux, and A. Telenti. Addressing the concerns of the Lacks family: Quantification of kin genomic privacy. In *CCS'13: Proc. of the 20th ACM Conf. on Computer and Communications Security*, pages 1141–1152, 2013.

[19] H. K. Im, E. R. Gamazon, D. L. Nicolae, and N. J. Cox. On sharing quantitative trait GWAS results in an era of multiple-omics data and the limits of genomic privacy. *American Journal of Human Genetics*, 90(4):591–598, 2012.

[20] A. Johnson and V. Shmatikov. Privacy-preserving data exploration in genome-wide association studies. In *KDD'13: Proc. of the 19th ACM Int'l Conf. on Knowledge Discovery and Data mining*, pages 1079–1087, 2013.

[21] M. Kantarcioglu, W. Jiang, Y. Liu, and B. Malin. A cryptographic approach to securely share and query genomic sequences. *IEEE Trans. on Information Technology in Biomedicine*, 12(5):606–617, 2008.

[22] M. Kayser and P. de Knijff. Improving human forensics through advances in genetics, genomics and molecular biology. *Nature Reviews Genetics*, 12(3):179–192, 2011.

[23] Z. Lin, A. B. Owen, and R. B. Altman. Genomic research and human subject privacy. *Science*, 305(5681):183, Jul 2004.

[24] F. Liu, F. van der Lijn, C. Schurmann, G. Zhu, M. M. Chakravarty, P. G. Hysi, A. Wollstein, O. Lao, M. de Bruijne, M. A. Ikram, et al. A genome-wide association study identifies five loci influencing facial morphology in europeans. *PLoS Genetics*, 8(9):e1002932, 2012.

[25] B. A. Malin and L. Sweeney. How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. *Journal of Biomedical Informatics*, 37(3):179–192, 2004.

[26] A. K. Manning, M.-F. Hivert, R. A. Scott, J. L. Grimsby, N. Bouatia-Naji, H. Chen, D. Rybin, C.-T. Liu, L. F. Bielak, I. Prokopenko, et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nature Genetics*, 44(6):659–669, 2012.

[27] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *SP'08: Proc. of the 29th IEEE Symp. on Security and Privacy*, pages 111–125, 2008.

[28] A. Narayanan and V. Shmatikov. De-anonymizing social networks. In *SP'09: Proc. of the 30th IEEE Symp. on Security and Privacy*, pages 173–187, 2009.

[29] http://www.ncbi.nlm.nih.gov/projects/SNP/.

[30] X.-l. Ou, J. Gao, H. Wang, H.-s. Wang, H.-l. Lu, and H.-y. Sun. Predicting human age with bloodstains by sjTREC quantification. *PloS ONE*, 7(8):e42412, 2012.

[31] A. Pollack. Building a face, and a case, on DNA. http://www.nytimes.com/2015/02/24/science/building-face-and-a-case-on-dna.html, Feb. 2015.

[32] R. Shokri, G. Theodorakopoulos, J.-Y. Le Boudec, and J.-P. Hubaux. Quantifying location privacy. In *SP'11: Proc. of the 32nd IEEE Symp. on Security and Privacy*, pages 247–262, 2011.

[33] M. Srivatsa and M. Hicks. Deanonymizing mobility traces: Using social network as a side-channel. In *CCS'12: Proc. of the 19th ACM Conf. on Computer and Communications Security*, pages 628–637, 2012.

[34] L. Sweeney, A. Abu, and J. Winn. Identifying participants in the personal genome project by name. 04/24/2013 2013.

[35] C. Troncoso, B. Gierlichs, B. Preneel, and I. Verbauwhede. Perfect matching disclosure attacks. In *PETS'08: Proc. of the 8th Privacy Enhancing Technologies Symp.*, pages 2–23,

2008.

[36] C. Uhler, A. Slavkovic, and S. E. Fienberg. Privacy-preserving data sharing for genome-wide association studies. *Journal of Privacy and Confidentiality*, 5(1), 2013.

[37] http://www.vox.com/2014/9/9/5975653/with-genetic-testing-i-gave-my-parents-the-gift-of-divorce-23andme. Last visited: Feb. 2015.

[38] S. Walsh, F. Liu, K. N. Ballantyne, M. van Oven, O. Lao, and M. Kayser. IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. *Forensic Science International: Genetics*, 5(3):170–180, 2011.

[39] R. Wang, Y. F. Li, X. Wang, H. Tang, and X. Zhou. Learning your identity and disease from research papers: information leaks in genome wide association study. *CCS'09: Proc. of the 16th ACM Conf. on Computer and Communications Security*, pages 534–544, 2009.

[40] X. Zhou, B. Peng, Y. F. Li, Y. Chen, H. Tang, and X. Wang. To release or not to release: Evaluating information leaks in aggregate human-genome data. *ESORICS'11: Proc. of the 16th European Conf. on Research in Computer Security*, pages 607–627, 2011.

[41] D. Zubakov, F. Liu, M. Van Zelm, J. Vermeulen, B. Oostra, C. Van Duijn, G. Driessen, J. Van Dongen, M. Kayser, and A. Langerak. Estimating human age from T-cell DNA rearrangements. *Current Biology*, 20(22):R970–R971, 2010.

Table 1. List of phenotypic traits and associated SNPs considered in the evaluation. Fields marked with a '*' are considered only in the supervised case.

High-level trait	Trait	Relevant SNP
Sex	Sex ("male", "female")	sexual chrom.
Skin color*	Type[10] ("II: white", etc.)	rs26722 rs1667394 rs16891982
Hair type	Curliness ("straight", "wavy", "curly", etc.)	rs7349332 rs11803731 rs17646946
Hair color	Blond ("yes", "no")	rs12821256 rs35264875
Eye color	Brown ("yes", "no")	rs916977 rs1129038 rs1800401 rs2238289 rs2240203 rs3935591 rs4778241 rs7183877 rs8028689 rs12593929
	Blue ("yes", "no")	rs1800407 rs7495174
Freckling*	Density ("light", etc.)	rs1042602
Ability to tan*	Ability ("yes", "moderate", "no")	rs1015362 rs2228479 rs1805009
Asthma*	Presence ("yes", "no")	rs5067 rs689465 rs2278206 rs2303067 rs7216389 rs11569562
Lactose tolerance*	Tolerance ("yes", "no")	rs182549 rs4988235 rs41380347 rs41525747
Blood type (e.g., "AB+")	Rhesus ("+", "-")	rs590787
	Has B ("yes", "no")	rs7853989
	Has O ("yes", "no")	rs505922 rs8176719
Earwax	Type ("dry", "wet", etc.)	rs17822931

Table 2. Sample probabilistic association models: $\mathbf{P}(\text{trait} \mid \text{SNP})$ for each relevant SNP-trait pair.

(a) Model for a quasi-deterministic association with two trait values.

	Blood type	
rs505922	**"O"**	**"Not O"**
CC	$1-\alpha$	α
CT	**0.5**	**0.5**
TT	α	$1-\alpha$

(b) Model for an association with three trait values (β_3 is such that $\beta_3+\beta_3^2+\beta_3^3=1$).

	Hair curliness		
rs11803731	**"straight"**	**"waivy"**	**"curly"**
AA	β_3	β_3^2	β_3^3
AT	**0.33**	**0.33**	**0.33**
TT	β_3^3	β_3^2	β_3