

Why Pseudonyms Don't Anonymize: A Computational Re-identification Analysis of Genomic Data Privacy Protection Systems

Bradley Malin
Data Privacy Laboratory, School of Computer Science
Carnegie Mellon University, Pittsburgh, Pennsylvania 15213-3890 USA
malin@cs.cmu.edu

Abstract

Objectives: In order to supply patient-derived genomic data for medical research purposes, care must be taken to protect the identities of the patients from unwanted intrusions. Recently, several protection techniques that employ the use of trusted third parties (TTPs), and other identity protecting schemas, have been proposed and deployed. The goal of this paper is to analyze the susceptibility of genomic data privacy protection methods based on such methods to known re-identification attacks.

Methods: Though advocates of data de-identification and pseudonymization have the best of intentions, oftentimes, these methods fail to preserve anonymity. Methods that rely solely TTPs do not guard against various re-identification methods, which can link apparently anonymous, or pseudonymous, genomic data to named individuals. The re-identification attacks we analyze exploit various types of information that leaked by genomic data that can be used to infer identifying features. These attacks include genotype-phenotype inference attacks, trail attacks which exploit location visit patterns of patients in a distributed environment, and family structure based attacks.

Results: While susceptibility varies, we find that each of the protection methods studied is deficient in their protection against re-identification. In certain instances the protection schema itself, such as singly-encrypted pseudonymization, can be leveraged to compromise privacy even further than simple de-identification permits. In order to facilitate the future development of privacy protection methods, we provide a susceptibility comparison of the methods.

Conclusion: This work illustrates the danger of blindly adopting identity protection methods for genomic data. Future methods must account for inferences that can be leaked from the data itself and the environment into which the data is being released in order to provide guarantees of privacy. While the protection methods reviewed in this paper provide a base for future protection strategies, our analyses provide guideposts for the development of provable privacy protecting methods.

Keywords: Privacy, anonymity, pseudonymization, trusted third parties, genomics, databases

1. Introduction

The genomic data of an individual is increasingly being collected, stored, and shared for various health related purposes. In both the research and clinical environment, genomic data provides opportunities for health care that until recently were severely limited or nonexistent. As a diagnostic tester for certain diseases, such as Phenylketonuria, early confirmation can initiate life-saving treatment, raise the standard of living, and help facilitate in family planning decisions. Beyond gross diagnostics, there is gathering evidence that suggests variation in our genome influences our body's ability to process drugs and our susceptibility to disease. However, with informative health-related data comes highly sensitive and personal information. Many people fear that information gleaned from their genomic data will be misused, abused, to influence their employment and insurance status, or simply cause social stigma [6, 9, 13, 20]. For individuals afflicted with a particular disease, such as Huntington's disease, diagnostic confirmation provides little hope or comfort because no cure or proven treatments exist. Moreover, an individual's genomic data, unlike most clinical health information, retains specific information on, and provides relationships about, related family members. Given the sensitivity of genomic data, there are many social pressures to protect the privacy of an individual's genomic status.

In addition to social pressures, there exist legal mechanisms for protecting genomic data privacy. In the United States, the adoption of the Privacy Rule of the Health Insurance Portability and Accountability Act [8], along with various state laws, mandate that the sharing of patient-specific data, including genome-based data, must protect the a patient's identity when their data is shared from the original source of collection. Failure to comply will result in legal action taken against the data sharer that can include revocation of government funding, fines, and imprisonment. Legal measures have been enacted outside of the United States as well. In the European Union,

the Data Protection Act (DPA) of 1998 imparts strict regulations on the processing of personal data, of which genomic data is a part of [7]. In order to process identifiable genomic data, a data subject's explicit consent must be secured before the processing of their data can begin. Yet, according to the DPA, if it can be shown that the data does not relate to a living individual, then the regulations are substantially relaxed. Thus, without proper guarantees of anonymity, not only will patients be less willing to provide data, but many data collectors will be unwilling to share genomic data for new or continuing collaborative research projects. In recognition of the problem, the protection of privacy for genomic data is considered a major challenge for the biomedical research community [1, 22].

Over the past several years, a variety of techniques have been proposed to protect the identity of an individual whose genomic data is shared. Several of the more sophisticated techniques advocate the use of pseudonyms to protect privacy [11, 18]. In general terms, pseudonymization converts the explicitly identifying features of an individual, such as name or social security number, into an encrypted or random value. The newly created value is referred to as a pseudonym. Advocates of such techniques claim that pseudonymization sufficiently protects the identity of the individual to whom the genomic data corresponds. At a glance, such claims appear to be true. How can one learn the identity of a pseudonymized genomic data sample, when there is no registrar linking pseudonyms to identity? Unless an adversary deduces the encryption keys via cryptanalysis, or breaks into an encrypter's computer and steals the key, the adversary should not be able to determine the identities of the pseudonyms.

The claim that pseudonyms protect privacy is fundamentally flawed for a variety of reasons. One reason is that discussions about the protection capabilities of a pseudonymizing technique are provided under particular assumptions about the data sharing environment. For example, a

pseudonymization schema for the protection of identities in data released from a single institution can fail when multiple institutions are releasing data. Trail re-identification algorithms [17], which we review in this paper, have been shown to leverage unique patterns in location-visit patterns to link seemingly anonymous genomic data to named individuals. There exist additional flaws and, subsequently, additional methods to re-identify genomic data.

This paper is organized a series of protection method analyses. In the next section, several published identity protection schemas for genomic data are reviewed. Following each method, a re-identification attack, which the method is susceptible to, is introduced. Since particular methods are susceptible to multiple methods, in Section 3 a high-level susceptibility analysis of each protection method for the presented re-identification attacks is presented. Finally, the need for research into formal anonymity protection schemas and how these analyses can help in the design of new protection methods is discussed.

2. Protection and Re-identification Methods

In this section we analyze several published identity protection methods for genomic data. For each method we review the protection schema and analyze the susceptibility of protected data to various re-identification techniques. Specific re-identification techniques are reviewed and discussed with respect to protection methods as they are introduced. For this analysis, terminology is based upon relational database theory. A table $\tau\{A_1, A_2, \dots, A_n\}$ refers to a table with the attributes $A = \{A_1, A_2, \dots, A_n\}$, where each attribute is a semantically-defined category. Each row of a table is referred to as a tuple $t[a_1, a_2, \dots, a_n]$ and corresponds to specific values $a_1 \in A_1, a_2 \in A_2, \dots, a_n \in A_n$. When table τ is partitioned into identifiable and de-identified subtables, the table containing identifiable data is referred to as τ^+ and the table containing de-

identified data is referred to as τ^+ . For example, in Figure 1, the table $\tau\{Name, Social Security Number, Pseudonym, DNA\}$ is partitioned into $\tau^+\{Name, Social Security Number\}$ and $\tau^-\{Pseudonym, DNA\}$.

τ			
τ^+		τ^-	
Name	SSN	Pseudonym	DNA
Bob	901231232	SA9212OK19	cttg...a
Kate	874017412	AS09D8LK1J	atcg...t
John	213732120	D8A79AD133	acag...t
Mary	521230138	ASSD834MS1	accg...a

Fig. 1. Table $\tau\{Name, SSN, Pseudonym, DNA\}$ partitioned into a table with identifiable data $\tau\{Name, SSN\}$ and de-identified data $\{Pseudonym, DNA\}$

2.1. deCODE

The first genomic data protection model we study was introduced by deCODE Genetics, Inc. [11] in 2000. The model consists of two parts, each of which uses a trusted third party (TTP) intermediary. The first corresponds to how appropriate research subjects are discovered in a de-identified manner. The second part entails how actual data is submitted to deCODE and used in-house for research purposes. Each is susceptible to different re-identification attacks. For this analysis, and brevity, we concentrate on how re-identification can occur during the research subject discovery process.

2.1.1 deCODE Trusted Third Party Model – Potential Patient Set Construction

The first part of deCODE’s research model is to determine an appropriate set of research subjects. Fig 2. provides an overview of this process, while the following text provides some more specific features of the procedure. According to the deCODE model, data collection is

handled on a research project specific basis. As such, the process is initiated by deCODE researchers. The researchers communicate a specific disease of interest to physicians who attend to the general patient population. The physicians create and send a population-based list $L\{Name, Social\ Security\ Number, \{Additional\ Identifying\ Data\}, Disease\}$ of patients with the disease to the Data Protection Commission (DPC) of Iceland, where “*additional identifying data*” pertains to a set of useful demographic attributes for the identification of individuals. Upon receiving this list, the DPC removes all explicitly identifying information, except for the Social Security Number (SSN). The social security number is encrypted with a reversible encryption function f into an alphabet-derived code $f(SSN)$.

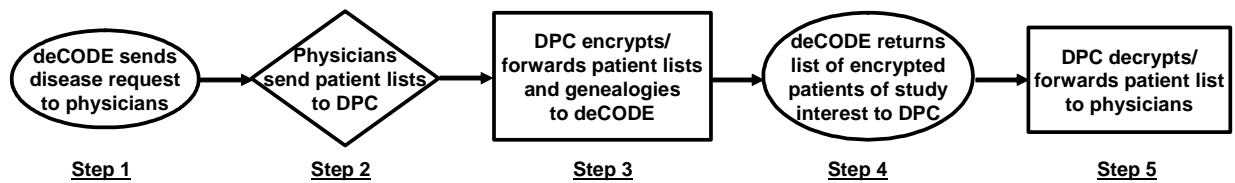


Fig. 2. Determining the set of relevant patients. The shape of the step denotes, which entity acts on the data: oval = deCODE, diamond = physicians, rectangle = Icelandic Data Protection Commission.

The encrypted list $L'\{f(SSN), Disease\}$ is then sent onto deCODE. At deCODE, the L' is fed through a computerized population-based genealogy, which has been previously encrypted by the DPC with encryption function f and linked to patient medical information. From the genealogy, deCODE determines a new encrypted list $N\{f(SSN)\}$ of individuals that they would like to gather genomic data from. Subsequently, list N is sent to the DPC, who decrypts the names and sends the list $N'\{name, SSN\}$ to the appropriate attending physicians for contacting their patients. By acting as the intermediary with full encryption and decryption capabilities, the DPC functions as a completely trusted third party (TTP).

2.1.2 Family-Structure Attacks

There are several re-identification techniques that can be employed to re-identify data protected under the deCODE model. We will go over an extended set of techniques in Section 3; here we focus on a particular attack, which we term the *family structure attack*.

deCODE considers the use of historical and genealogical repositories to be one of its most powerful techniques for discovering interesting patterns and useful patients. It should be apparent that pseudonymizing the names of individuals does not obscure any of the genealogical structures. Genealogies, which are rich in depth and structure, permit the construction of large sets of familial relationships. In theory, such families can have many configurations and variation. However, it is the massive quantity of family structure configurations that permit a re-identification attack.

Let us analyze the complexity and number of family structures that one can discern. As a base case, consider one of the simplest blood-related family structures, as shown in Fig. 2., of 2 parents and 1 child. Since there must exist a man and woman to produce a child, the only variable is the gender of the child. For this analysis, we use the variable V_i , to represent the ambiguous variable of gender for child i , which can undertake the values of male (M) or female (F).

In the family structure, there are 2 variants on the family structure: $[M_1, F_1, V_1=M]$ and $[M_1, F_1, V_1=F]$.

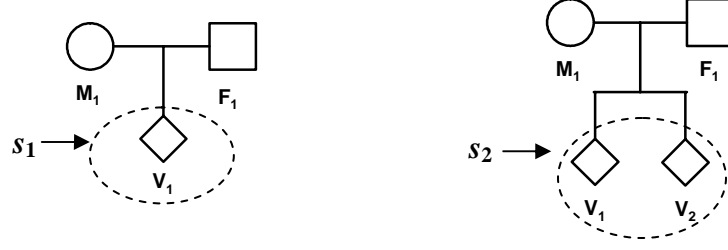


Fig. 2. Pedigrees of nuclear family with left) one child (s_1) and right) two children (s_2). Right) Circle = male (M), square = female (F), and diamond = variable (V) gender.

Since we are considering not only pedigrees, but diseases status, we must incorporate disease status into the problem. Let us consider that disease status D is a Boolean variable, such that an individual is either diagnosed with D or is not. In this aspect, all three variables M_1 , F_1 , and V_1 are independent of each other and there exist $|\{M_1, D\}| * |\{F_1, D\}| * |\{V_1, D\}| = 2 * 2 * 4 = 16$ possible variants on what we term the *family-disease structure*. Now, let us expand the simple family to include two children. In this family, V_1 and V_2 are not independent, since $[V_1=F, V_2=M]$ is equivalent to $[V_1=M, V_2=F]$. There are 3 variants on this family structure: $[M_1, F_1, V_1=M, V_2=M]$, $[M_1, F_1, V_1=M, V_2=F]$, and $[M_1, F_1, V_1=F, V_2=F]$. And when disease status is factored in, there exist $|\{M_1, D\}| * |\{F_1, D\}| * |\{V_1, V_2, D, D\}| = 2 * 2 * 10 = 40$ variants on the family-disease structure.

We derive a recursive relationship for calculating the number of sibling-disease structures for an arbitrary number of siblings as follows. Let s_n be the number of variants for a sibling-disease structure consisting of n children. First, it should be obvious that $s_0=1$, since there exists only one type of family without any children. Now, $s_1= 4$ as we demonstrated above for the lone sibling structure of $\{V_1, D\}$. We define the recursion as:

$$s_0 = 1, \quad s_n = s_{n-1} + \frac{(n+1)(n+2)}{2}$$

From a computational standpoint, the above recursive relation can be simplified into a relatively simple series, which can be solved via induction (see Appendix A) The resulting closed form expression is:

$$s_n = \sum_{i=0}^n (n-i+1)*(i+1) = \frac{n^3 + 6n^2 + 11n + 6}{6} \quad \text{Eq. 1}$$

Eq. 1 provides a direct way of computing the number of variants for a sibling-disease structure of n siblings. When computing the number of variants for a nuclear family-disease structure with s_n , the parents of the siblings must be factored in. Let us call this number y_n . In a naïve scenario, the disease status of the children is independent of the parents. As presented above, this $|\{M_1, D\}|*|\{F_1, D\}| = 4$, and thus $y_n = 4*s_n$. From y_n and n we can compute the total number of individuals that are covered by the variants of the family-disease structure, which is $y_n*(n+2)$. Actual numbers for such family-disease structures are presented in Table 1 for families of up to 6 children.

Table 1. Uniqueness of family structures and the subsequent number of re-identifications possible for a simple nuclear family.

# of Children (n)	# of Sibling Structure Variants	# of Sibling-Disease Structure Variants (s_n)	# of Nuclear Family-Disease Structure Variants (y_n)	Max # of Individuals Re-identifiable $(n+2)*(y_n)$
0	1	1	4	8
1	2	4	16	48
2	3	10	40	160
3	4	20	80	400
4	5	35	140	840
5	6	56	224	1568
6	7	84	336	2688

Since genealogical data is mainly of interest, pedigrees will almost always be much more robust than a simple nuclear family. Thus, consider a more complex scenario with a family that consists of three generations of family members (i.e. grandparents, parents, and children). An extended family hinged around one set of parents is presented in Fig. 3. There are 4

grandparents, 2 parents, 2 siblings of the parents, and 2 children of the parents. The image of Figure 3 provides the generalized family structure with 2 children for each set of parents. The right image of Fig. 3 provides a specific variant of the family-disease structure.

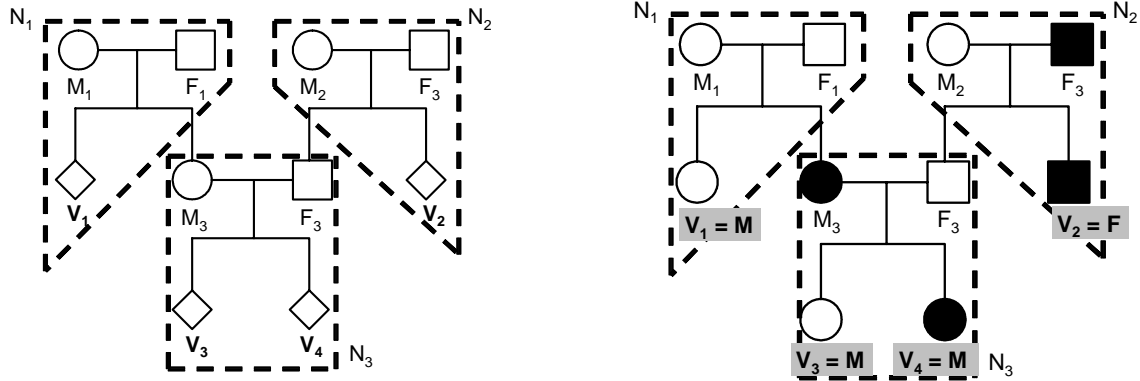


Fig. 3. Left) Simple extended family pedigree hinged on parents M_3 and F_3 . Independent nuclear families are outlined and denoted N_1 , N_2 , and N_3 . Right) A specific variant of the family-disease structure.

Based on this structure, all males and female are fixed because they must be alive for V_3 and V_4 to exist. Notice that there are three independent sets of sibling structures $\{V_1, M_3\}$, $\{V_2, F_3\}$, and $\{V_3, V_4\}$, and since M_3 and F_3 are fixed, we can reduce the first two sets to $\{V_1\}$ and $\{V_2\}$, respectively. Based on these sets, there can exist $|V_1|*|V_2|*|\{V_3, V_4\}|$ distinct familial structures, which in numbers is $2*2*3 = 12$ distinct types families. Factoring disease status in, there are $s_1*s_1*s_2 = 4*4*10 = 40$. Considering the disease status of all family members, the number of distinct family-disease structures is equivalent to $|N_1|*|N_2|*|N_3| = y_1*y_1*y_2 = 16*16*40 = 10240$ distinct family structures.

Note that this number accounts only for a very particular family structure. In the event that V_1 and V_2 each have a 2 child family, then the number of variants of the family disease structures explodes to an order of 10^6 family-disease structures. Factoring in the various types of sibling structures that exist, there exist $(y_0+y_1+y_2)^3$ or over a 200,000 variants on the family-disease structures hinged around two parents as shown in Figure 2, where the number of children varies

from 1 to 3 for each of the grandparents and 0 to 3 for the core parents. From this analysis, it should be apparent that it does not take much to make a genealogy unique.

Obviously, not all disease-family structures, and the variants of such, will be realized in a population, and certain variants will be more probable than others. However, a magnitude of 10^6 remains a daunting number, considering that there are approximately 10 individuals in this family structure, and thus on the order of 10^7 individuals in consideration. The number of family disease structures is even larger, considering that this analysis does not account for additional features, such as the fact that certain family members may be deceased (another identifying feature which can be communicated in the pedigree), and that supplementary medical or genomic features can be included in the health information [Cancer]. The latter is should be especially noted, since many polygenic trait studies are interested in learning which factors are the most influential in disease severity or occurrence.

The ability to determine unique pseudonymized family structures is one part of the re-identification susceptibility. Alone, unique family structures do not reveal identities. Thus, in addition to unique family structures, we need identifiable information to link our family structures to. However, such identifiable information is publicly available in the form of various genealogical databases available both offline on CD-ROMs and on the World Wide Web. For example, public genealogical records on Icelanders, reside in many popular publicly available databases, including Ancestry.com, Infospace.com, RootsWeb.com, GeneaNet.com, FamilySearch.org, and Genealogy.com.¹ From such data, it is not difficult to construct identifiable family structures. And with such information in hand, an adversary can link disease labeled family structures to named individuals.

¹ At the time of writing, the website <http://www.rat.de/kuijsten/navigator/iceland/> provided a larger number of Icelandic genealogical resources.

2.1.3 Trusted Third Party Model – Patient Sample Submission

After a patient is contacted by their physician, they can choose whether or not to participate in the deCODE research study. The set of patients that opt-in donate blood samples at a facility run by the DPC, where each sample is labeled with a “sample number” (AN). The AN and SSN of the individual is then entered into a computer database. Subsequently, the DPC encrypts the SSN, with their encryption function f , into $f(SSN)$ and forwards both the AN-labeled samples and the encrypted list of participating subjects $P\{AN, f(SSN), Sample, Disease\}$ to deCODE. At deCODE, each AN is associated with a new in-house number IN, such that deCODE maintains the list $P'\{IN, AN, f(SSN), Sample, Disease\}$. Though the AN-IN relationship is known at deCODE, AN's are withheld from the majority of laboratory and deCODE employees. Rather, the laboratory researchers at deCODE work with $P''\{IN, Sample, Disease\}$.

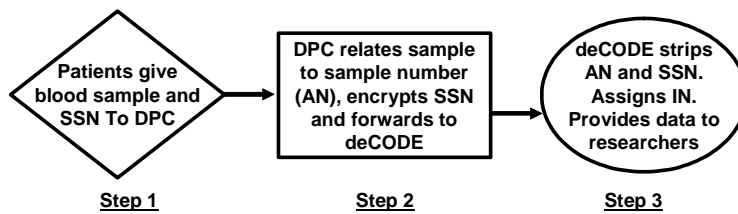


Fig. 4. Participating patient data de-identification. The shape of the step denotes, which entity acts on the data: oval = deCODE, diamond = physicians, rectangle = Icelandic Data Protection Commission.

2.2. Gent

The second model considered was introduced by researchers from the University of Gent [18]. In this work, both a batch and interactive protection model for protecting the privacy of genomic data are outlined. For this research, our analysis is limited to the batch model.

Akin to deCODE, the Gent model employs a trusted third party (TTP); however, the TTP is given restricted access to the patient-specific information. In the deCODE model, trust is necessary on the part of both the physicians and the patients, since they supply raw data to the DPC, who encrypts the data before passing it onto deCODE. In the Gent model, the groups that submit DNA for secondary usage do not have complete trust in the TTP. More specifically, they believe that the TTP should not be permitted to see the identities of the individuals whose DNA has been submitted. When a third party does not have complete access to data, we say that it is a semi-trusted third party (sTTP).

2.2.1 *Semi-Trusted Third Party Model*

Under the Gent model, a set of data holders, such as a set of physicians or researchers at a particular institution that have collected genomic data from a set of patients, transfer data to a central repository maintained by a third party. First the data holders collect data on their patients and construct a list of identified individuals and their genomic data $L\{Identity, DNA\}$. Instead of handing over the raw data to the third party, the data holders apply a public-key encryption function f to the *Identity* attribute and convert L into $L'\{f(Identity), DNA\}$. This encrypted list is passed onto the sTTP, who then applies their own public-key encryption function g to $f(Identity)$ in order to create a doubly-encrypted list $L''\{g(f(Identity)), DNA\}$. In addition, the sTTP can act as a data broker for multiple data holders. Thus, the data holder maintains a set of lists, $A\{g(f_A(Identity)), DNA\}$, $B\{g(f_B(Identity)), DNA\}$, ..., $Z\{g(f_Z(Identity)), DNA\}$ for locations A , B ... Z , each of which uses their own encryption functions and keys.

When a new researcher requests sTTP for data, sTTP supplies the appropriate set of doubly-encrypted lists. In the event that researchers need additional data on the subjects in the supplied

lists, such as from location A , then the researchers send a request onto sTTP with a sublist of individuals $A'\{g(f_A(Identity))\}$. In turn, the sTTP decrypts and sends the single-encrypted pairs $A'\{f_A(Identity)\}$ onto location A for additional data.

2.2.2 Trail Attacks

One of the main reasons why data requesting researchers are supplied with doubly-encrypted pseudonyms is to prevent adversaries from employing a dictionary attack for re-identification purposes. Despite resilience to a direct dictionary attack, which is described more in-depth in Section 3, the Gent model does not guarantee identity protection. One of the main reasons why this method fails is that it is susceptible to what is known as a *trail* attack. The trail attack works as follows. Consider an environment where there exists a set of locations L , such as a set of hospitals, and a set of data subjects S , such as a set of patients. At each location $l \in L$ that a patient visits, l has the ability to collect multiple types of information, such as clinical and genomic data. To protect privacy when data is released, each hospital releases data in a partitioned manner, such that identified data and de-identified data are released separately. The first table released, $\tau^+(demographic\ information, clinical\ information)$, where *demographic information* contains identifiable data. Oftentimes, patient clinical and/or discharge data is released in an identified manner or a de-identified manner that can be re-identified to named individuals through linkage with public records [21]. Therefore, let it be the case that identified clinical information is available on the set of patients. The second table released, $\tau^-(DNA)$, consists of as one partition and a list of genomic data samples as another.

As stated above, by the Gent model, the genomic data list consists of encrypted identifiers. Thus, if an adversary was to attempt a linkage of the genomic data with the patient list for any

particular location, then no re-identifications can occur. However, it is apparent in many environments, including the Gent model, that many locations function and release data independently. This aspect is detrimental to privacy since both genomic and explicitly-identifiable data consist of static information. Thus, if an adversary retrieves data from a set of locations, he can create two tracks of data. The first track consists of the set of locations that a genomic data sample was left behind at, or the trail of the data. The second track consists of the set of locations that an identity visited. Based on trail patterns in these two tracks, it is possible to link the genomic data trail left behind by an individual to the trail of explicitly identifiable data. Fig. 5. provides an example of how identified and DNA tracks can be constructed.

		l_1	
τ^+		τ^-	
<i>Name</i>		<i>Pseudonym</i>	<i>DNA</i>
John		1G09JU3R	acag...t
Mary		F4P02SD4	accg...a

		l_2	
τ^+		τ^-	
<i>Name</i>		<i>Pseudonym</i>	<i>DNA</i>
John		4FG5097H	acag...t
Bob		U89KM32J	cttg...a

		l_3	
τ^+		τ^-	
<i>Name</i>		<i>Pseudonym</i>	<i>DNA</i>
Kate		AOEHA120	atcg...t
Bob		1X3C5VK4	cttg...a
Mary			

Identified Track			
<i>Name</i>	l_1	l_2	l_3
John	1	1	0
Mary	1	0	1
Bob	0	1	1
Kate	0	0	1

DNA Track				
DNA	Pseudonyms	l_1	l_2	l_3
acag...t	1G09JU3R 4FG5097H	1	1	0
accg...a	F4P02SD4	1	0	0
cttg...a	U89KM32J 1X3C5VK4	0	1	1
atcg...t	AOEHA120	0	0	1

Figure 5. Left) Data releases made from three locations l_1, l_2, l_3 . Right) Resulting Identified and DNA tracks created.

Given identified and DNA tracks, formal algorithms to perform trail re-identification have been developed [15, 17]. These methods carry out re-identification in a variety of circumstances, including when genomic data and identity trails are exactly the same, as shown in the left of Fig. 5., when certain data is missing, as shown in the right of Fig. 5., and even when a patient leaves behind multiple samples (e.g. sequence of a gene vs. SNP data) [16].

2.3. De-identification

The third model we examine has been employed by a variety of groups. It exists in many environments, including the sharing of epidemiological and genetic data [4, 23], the construction of human mutation databases, and research with disease registries [10]. This method entails the simple de-identification of the data being studied.

2.3.1 *Random IDs and De-identification*

Advocates of the de-identification model for DNA protection assume that the identity of DNA data is protected if explicit identifiers are removed or generalized. Under a de-identification model, the identifying information of an individual is obscured completely or generalized before being released. For example, instead of releasing the date of birth, the age of the individual would be released. In many situations, a unique identifier is assigned to a patient for linkage purposes. In the model designed for the Utah Resource for Genetic and Epidemiologic Research (RGE), the identifier is a random number, which is generated for each subject. In this case, the random number functions as a pseudonym.

2.3.2 *High-Level Inference Attacks*

The *high-level linkage attack* is an attack strategy that utilizes domain knowledge. In prior research, this method was demonstrated with a direct linkage strategy. It was proven that the removal of a set of attributes from one list removal of direct identifiers, such as date of birth, do not guarantee anonymity [21]. The high-level linkage attack proceeds as follows. Given two tables $X\{A_{x1}, A_{x2}, \dots, A_{xm}\}$ and $Y\{A_{y1}, A_{y2}, \dots, A_{ym}\}$, we construct a set of relations XR_Y between

the two tables. For example, in the data protection method of the RGE, released genomic data may be accompanied by de-identified demographic information. In an attempt to prevent direct linkage, certain demographic information, such as date of birth may be generalized simply to age of the subject. Though the generalization of date of birth may appear to protect data, it is not guaranteed. For example, let the two tables be *Health*{*name, address, birthdate, gender, zip code, hospital visit date, diagnosis, treatment*} and *DNA*{*age, gender, hospital visit date, DNA*}. The obvious set of attribute relationships that can be extracted are {<*birthdate, age*>, <*gender, gender*>, <*hospital visit date, hospital visit date*>}. For each tuple *h* in *Health*, we can check the number of tuples in *DNA* that *h* can be related to. If the number of tuples is 1, and we have sufficient belief that this is the only entity in the general population that *h* could be related to, then a re-identification is revealed.

The relation set that was constructed can be expanded if relationships between clinical and genomic data exist. In fact, in previous research, we demonstrated that such knowledge can be extracted. We discovered that there exist a significant number of diseases for which a mutation in the genome is directly related to a standard International Classification of Disease Code - version 9 (ICD-9) [14]. We recently performed a non-exhaustive literature review, and discovered that there exist at least 40 ICD-9 codes that can be related to 37 DNA-mutation causing diseases. This list we provide in Table 2. In addition, the pharmacogenomics community continues to discover relationships between the variation in an individual's genome and the ability to process drugs and treatments [1]. Given such domain knowledge, it is easy to expand the relation set to include such relations as {<*diagnosis, DNA*>, <*treatment, DNA*>}.

Table 2. ICD-9 codes that can be inferred from mutations in a patient's genomic data.

Disease in Medical Release Data	ICD-9 Code	Known Gene(s)
Adrenoleukodystrophy	3300	ALD

Amyotrophic Lateral Sclerosis (ALS)	33520	SOD1, ALS2, ALS4, ALS5
Burkitt's Lymphoma	2002	MYC
Chronic Myeloid Leukemia	2051, 20510, 20511	BCR, ABL
Cystic Fibrosis	27700, 27701, V181, V776	CFTR, CFM1
Duchenne's Muscular Dystrophy (paralysis)	33522	DMD
Ellis-van Creveld (chondroectodermal dysplasia)	75655	EVD
Essential Tremor (idiopathic)	3331	ETM1 (FET1), ETM2
Familial Mediterranean Fever (amyloidosis)	2773	FMF
Fragile X	75983	FMR1
Friedrich's Ataxia	3340	FRDA
Galactosemia	2711	GALT
Gaucher's disease (cerebroside lipidosis)	2727, 3302	GBA
Hemophilia Type A	2860	HEMA
Hereditary Hemorrhagic Telangiectasia	4480	HHT
Huntington's Chorea	3334	HD
Hyperphenylalaninemia (Phenylketonuria)	2701	PAH
Immunodeficiency with hyper-Igm (HIM)	27905	TNFSF5
Kugelberg-Welander	33511	SMN/NAIP region
Machado-Joseph Disease (Spinocerebellar Ataxia 3)	3348	MJD
Marfan Syndrome	75982	FBN1
Menkes Syndrome	75989	ATP7A
Methemoglobinemia	2897	HBB, HBA1, DIA1
Myotonic dystrophy	3592	DM
Pendred's syndrome	243	PDS
Prader-Willi Syndrome	75981	SNRPN
Refsum's Disease	3563	PAHX
Sickle Cell Anemia	28260	HBB
Spinocerebellar ataxia/atrophy	3349	SCA1
Tangier disease	2725	ABC1
Tay-Sachs	3301	HEXA
Tuberous Sclerosis (Pringle's disease)	7595	TSC1, TSC2
Vitelliform Macular Dystrophy (Best Disease)	36276	VMD2
von Hippel-Lindau (Angiomatosis Retinocerebellosa)	7596	VHL
Werner's disease or syndrome	2598	WRN
Werdnig-Hoffmann disease	3350	SMA1
Wilson's Disease	2751	ATP7B

2.3.3 Low-Level Inference Attacks

In the high-level attack, we utilize information that can be directly extracted from a single tuple of information. However, when more robust clinical information is available, additional information about an individual's genomic data can be inferred. When relational information is learned from multiple sources of information or multiple tuples in the same table, we call it a *low-level inference attack*. For example, consider Huntington's disease, which is caused by a CAG repeat mutation in the HD gene. Research has shown that there exists a relationship between the age of onset of the disease and the number of CAG trinucleotide repeats [2, 3]. Unlike the relationships between gene mutations and clinical codes constructed for the high level

analysis, the age of onset is not standard medical information included in an individual's medical information. Yet, we discovered that, given longitudinal medical information on an individual, tight bounds for the age of onset (*i.e.* within 3 years of age) of the patients could be constructed [16]. With this knowledge, one can generate and append a distribution of the expected age of onset to both the medical and DNA data. Each distribution is estimated from its respective information.

An additional feature of the low-level inference attack for genomic data is that it can become more powerful with time. Since the goal of genomic medicine is to elicit the relationships between genomic data and clinical phenotype, the number of relations, and specificity of such, increase with advances in basic medical research. Consider the Huntington's disease example. In the early stages of Huntington's disease research it was determined that a mutation in chromosome 4 was responsible for the manifestation of the disease in a patient. At that time, the only clinical information that a genome could reveal was Boolean, either an individual would be diagnosed with the disease or they would not. Yet as time progressed, and research into the disease became more advanced, it was found that the disease was caused by a CAG repeat mutation and that this mutation correlated with the disease's age of onset. At first, a gross correlation was determined, such that the juvenile form of the disorder could be distinguished from the adult form. After continued research, it was shown there exists a strong correlation between the general age of onset of the disease and CAG repeat size. This pattern of increased granularity is due to increased sophistication in the understanding of genotype-phenotype correlations. Moreover, such relationships are not limited to health related information, but to demographic information. The goal of the human genome diversity project and genomic anthropology is to determine the relationships between genomic variation and ethnicity.

2.4. Denominalization

2.4.1 Random and Family Coding

The model proposed by Gaudet et al. [10], which we will refer to as the Quebec model, approaches the problem of identity protection from a partitioning perspective. The technique is referred as “denominalization” or the separation of features corresponding to identity. The technique works as follows. Each patient is assigned two numbers. The first is a random number, while the second is a family-based code. The random number is used to manage the clinical and biological samples corresponding to the individual. In the protocol description, it is claimed that different levels of anonymity can be achieved through the suppression of various identifiers. Fully anonymous biological samples are those that are stripped of all identifiers.

The family based code is a numerical code that is subdivided into five parts or cells. The first cell corresponds to clinical aspects of an individual, while the latter four cells refer to 1) family number, 2) relation to family member, 3) child of which marriage (i.e. if parents have had multiple marriages), and 4) relation as a sibling, respectively. In addition, an individual may have a family based code for multiple families, with a connecting code relating the information. The individual and family codes are managed independently. Information corresponding to family codes is then released in controlled manner through the withholding of particular cells.

2.4.2 Pseudonyms as an Anonymity Threat

The denominalization model is susceptible to both the family-structure and inference attacks described above. However, this method is susceptible to another attack, one that is dependent on the pseudonym and coding strategies that denominalization utilizes.

As demonstrated above, methods that mask the explicit identity of genomic data through such methods as pseudonyms or de-identification offer no protection against particular types of re-identification. Moreover, blind faith in cryptographic or recoding methods to protect anonymity can provide the basis for further erosion of patient privacy, beyond that of a susceptibility to re-identification. Consider a set of hospitals H , where each hospital $h \in H$ releases tables τ_h^+ and τ_h^- with attributes $A_h^+ = \{name, date\ of\ birth, gender, zip\ code, clinical\ data\}$ and $A_h^- = \{pseudonym_h, DNA\}$. The attribute $pseudonym_h$ is generated through a reversible encryption function f_h , such as public-key encryption $f_h(Identity, key_h) = pseudonym_h$, where $Identity$ is a tuple of patient information $[name, date\ of\ birth, gender, zip\ code]$. Richard (*a.k.a.* Richard the Re-identifier) is an unscrupulous researcher who wants to re-identify as much data about the patient population as possible. So, Richard uses a trail re-identification attack to re-identify some of the patients released from a set of data releasing locations. Upon re-identification, Richard can construct a table with the attributes $\{name, date\ of\ birth, gender, zip\ code, pseudonym_1, pseudonym_2, \dots, pseudonym_H\}$, where $pseudonym_x$ is the pseudonym that hospital x uses for the identity of the patient. Thus, Richard has achieved his goal of re-identifying the protected genomic data.

A modified version of the dictionary attack can be used to exploit information released under the Quebec model. First, recall the family-structure attack described above. Given sufficient information to reconstruct and re-identify a certain amount of familial information, the recoding of familial relations can reveal additional information that may or may not have been learned in the family-structure attack, such as temporal information in the genealogy. For example, when a family has multiple children, the fifth cell of the family code, denotes what order of birth a

sibling is. Moreover, under the coding schema, this information is distinguishable for males, where the system uses even numbers, and females, where odd numbers are employed.

Furthermore, with this dictionary information in hand, there are additional malicious acts that Richard is now capable of conducting. When location x has sufficient confidence in the pseudonymization process to uphold anonymity, employ $pseudonym_x$ is utilized for internal data linkage purposes. Though the Quebec model explicitly separates individual information from family information, there are other models that do use $pseudonym_x$ for multiple purposes. This means that if Richard requests additional information about a set of individuals with pseudonyms (which he has already re-identified) from location x , he can learn additional information about patients that was meant to remain anonymous.

There is an additional problem though, which has to do with the susceptibility of the re-identified data to cryptanalysis. For each location x , Richard can construct a set of $\langle Identity, pseudonym_x \rangle$ pairs. Given a set large enough, Richard will be able to conduct a dictionary attack to learn the encrypting function and key for location x . With function and key in hand, Richard can perform a variety of malicious acts. First, since trail re-identification may not have re-identified the identities of all individuals, with Richard can now decrypt the additional pseudonyms that he failed to re-identify. The decrypted pseudonyms can then be directly linked to the clinical data. In addition, Richard may be able to pose as a confidant of location x and generate false data through the key. Data generated by Richard will appear to be real, since he can incorporate genuine demographic data to encrypt pseudonyms. In this manner, Richard could submit data to affect research project datasets or falsify certain patients' medical records.

3. Re-identification Susceptibility

In the previous section, it was shown that none of the genomic privacy protection methods are impervious to re-identification. This was presented from the perspective of a single re-identification attack being used against a given privacy protection method. However, though the susceptibility of each method was discussed for one method, does not suggest that the method provides protection against the other methods. In fact, we find this to be quite to the contrary. In this section, we examine the general re-identification susceptibility for each of the protection methods.

Table 3. Gross susceptibility of privacy protection models to re-identification.

	deCODE	Gent	Quebec	De-identification w/ Random IDs
Third Party	Full	Semi	N/A	N/A
Model	Trusted Third Party Encryption	Semi-Trusted Third Party Encryption	Denominalization and recoding	De-identification / Random ID
Family Structure Attack	Yes	No	Yes	Yes
Trail Attack	No	Yes	No	Yes
High-Level Inference Attack	Yes	Yes	Yes	Yes
Low-Level Inference Attack	No	Yes	Yes	Yes
Dictionary Attack	Yes	Yes	No	No

In Table 3, a side-by-side comparison of susceptibility of the protection models to known re-identification attacks is reported. This analysis is presented from a general point of view, such that either a technique is susceptible or it is not. We find that each of the protection models is susceptible to a minimum of three of the four re-identification attacks. Here, we discuss how each of the re-identification attacks fares against the protection models.

Family Structure Attack. The only model not susceptible to the family structure attack is the Gent semi-trusted third party model. Under this model, no familial relationships are considered in the genomic data. Under very specific cases, familial inferences may be possible, such as

haplotype analysis of DNA sequences to determine possible familial relations. However, without more confidence about whether or not related family members are in the dataset, such analysis could create false family structures and familial relations. Note that the goal of the Quebec model's denormalization strategy strives to prevent the family attack almost explicitly. It provides protections by separating the individual from the family and using a local recoding of the identity. Yet, once this information is studied in a genealogical setting, the protections are minimal. Similarly, both the deCODE and the RGE models reveal genealogical information. The deCODE model does so on large scale, since this is how subject recruit is performed. In contrast, the RGE model is more difficult to analyze. The RGE model is more difficult to analyze than the previous three. In general, as shown in Table 3, the RGE model is susceptible to all re-identification attacks. Yet, this may be somewhat deceiving. Since the RGE maintains a massive repository of diverse datasets, not all re-identification attacks can be performed on every dataset released by RGE. The analysis of re-identifiability for RGE released datasets is data dependent. Since RGE does have the ability to reveal genealogical information, and the only protection afforded to such data is de-identification and pseudonymization with random ID's, this model is susceptible to the family structure attack.

Trail Attack. To construct a trail attack, two criteria must be satisfied. The first requirement is that an individual's data can be distributed over multiple locations. The second requirement is that both the genomic and the identified data are available in a partitioned manner. Table 4 provides a characterization of which features the protection methods satisfy. The deCODE model does not satisfy the multiple location criteria. No location based information is not revealed, nor is necessary. In addition, the Quebec model is not susceptible. Under the current

version of the Quebec model, genomic data is collected at one location only. Yet, if this model was applied to a distributed information collecting and sharing environment, then the trail attack would be a feasible method of re-identification. The Gent model, as discussed above, does satisfy both criteria and is therefore susceptible. The RGE model is susceptible as well, since genomic data could be requested from multiple sources. The health-specific information could be either supplied directly as a separate source, or derived from various external resources, such as discharge information.

Table 4. Trail re-identification criteria.

Model	Multiple Locations	Partitioned Identified and DNA Data Available
deCODE	No	Yes
Gent	Yes	Yes
Quebec	No	Yes
RGE	Yes	Yes

High-Level Inference. This inference attack exploits the relationships that can be constructed between the genomic data and known demographic or clinical information. As such, all four protection methods are susceptible to the attack. However, it should be noted that though relationships between attributes in different datasets can be constructed, this does not guarantee that a re-identification will occur. Thus, when considering high-level inference attacks with genomic data by itself, as is the case with the Gent model, this attack is dependent on the specificity of the known relationships between genomic data and clinical phenotype. Yet, when demographic and geographic information may accompany the genomic data, such as in the RGE model, then care must be taken to determine how uniquely an individual’s features identify an individual.

Low-Level Inference. When more specific and longitudinal information is available, then this attack is applicable. The deCODE model does not permit this type of attack because the goal is for gene discovery, not the clinical applicability of variation nor the derivation of more specific diagnostic information. Note that this paper only concerns the patient recruitment deCODE model, and that this exemption from attack may not apply to other aspects of the deCODE research and development models. In contrast, such finer-grained genotype-phenotype research is of interest in the three other models studied. As such, these methods can leak relationships that while useful for research purposes and correlation studies, may allow for unique linkages to be constructed between identified and genomic data.

Dictionary Attack. With respect to the dictionary attack, the most susceptible type of model is that which uses a single pseudonymization function, where pseudonyms are derived from patient-specific information. Since the RGE model uses random ID's for pseudonyms, there is no way to run dictionary attack, regardless of the number of people re-identified through other means. Nonetheless, the deCODE and Gent models are both susceptible to dictionary attacks. In the deCODE model, this attack can be applied by end-users, or the researchers that request pseudonymized information. The genealogies studied by deCODE contain single-encrypted pseudonyms based on the social security numbers of the Icelandic population. The dictionary attack that the deCODE model is susceptible to is cryptographic in nature. Basically, as more and more people re-identified, the adversary can collect a set of SSN, pseudonym pairs. Given enough pairs, the adversary can learn the key of the pseudonymizing function. In contrast, the pseudonyms used in the Quebec model, as discussed above, are constructed from familial

relationship information. Since this coding structure is already known, no dictionary attack is necessary – rather a familial-structure attack can be performed directly with the pseudonyms.

Considering the Gent model, the dictionary attack can not be applied by requesters of information from the semi-trusted third party (sTTP). This is because the pseudonyms supplied to the researchers are doubly-encrypted. Though not random, it is almost impossible to discern the effects of the original sources pseudonymizing function from the sTTP's. However, a dictionary attack can be utilized by the sTTP itself. In the event that the sTTP is corrupt, it can leverage the fact it is receiving single-encrypted pseudonyms from each of the submitting sources.

4. Compounding Re-identification Attacks

As presented above, in many cases a genomic data privacy protection model is susceptible to multiple re-identification attacks. What is interesting to note though, is that many of the re-identification attacks reviewed in this paper can be used in combination to assemble more robust re-identification methods. As was demonstrated above, the family structure attack can be used in combination with a high level inference attack to construct more robust family structures or with a dictionary attack when additional information on the family is known. Moreover, an iterative process of alternating re-identification methods can be employed. Since different re-identification methods exploit different types of information, one could imagine using one re-identification method to re-identify a certain number of individuals in the population and then using a second re-identification method to re-identify individuals that could not be re-identified until certain confounding entities were removed from consideration. This process can continue until no more re-identifications are possible with the known methods.

5. Discussion

Based on the system analyses above, it can be concluded that pseudonymization and naïve de-identification strategies are not sufficient mechanisms for the protection of identities. Yet, this realization does not imply that pseudonyms and third party solutions are worthless in the pursuit of genomic data privacy protection. Rather, to an extent, these systems do provide certain privacy protections. First of all, pseudonyms serve as a first-order protector and deterrent. It is conceivable that an adversary, who approaches re-identification in a non-computational manner, will be deterred by the simple obscuring of explicitly identifiable information. Secondly, datasets devoid of linkage capabilities severely limit the types of research that can be performed. It is often the case where researchers may need to request additional information about a subject. From another point of view, a subject may wish to remove their data from a research study or find out information about how their data is being handled. In this respect, pseudonyms provide an extremely valuable service by accounting for future research, applications, and auditing capabilities that would be virtually impossible to handle without a linkage mechanism.

And yet, something must be done to protect the identities of the research subjects. This research is a call to arms for the biomedical community. Researchers must develop privacy protection methods that incorporate guarantees about the protections that they afford. New methods must account for multiple environments of data sharing as well as the type of inferences that can be gleaned from the shared data itself. These methods must be developed in a more scientific and logical manner, with formal proofs about the protection capabilities and limitations afforded by the specific method. Though proofs may be difficult to derive in the face of uncertainties about the sharing environment, especially when the data itself holds latent knowledge to be learned at a

later point in time, researchers can validate their approaches experimentally against known re-identification attacks, such as those discussed above.

On the flipside though, researchers should not remain content with their proofs and experiments. New re-identification attacks will be developed by those in the academic community, and the adversaries that reside outside of the public realm. As such, researchers must continue to innovate and develop new methods re-identification for testing their protection techniques. These methods may new types of inferential or location-based techniques or completely new models that have yet to be discovered. Regardless, without the development of new protection and re-identification methods, researchers will continue to rely upon untested and possibly dangerous methods of privacy protection. The development of new identity protection strategies is paramount for continued data sharing and innovative research studies.

6. Conclusion

This research provides an analysis of the re-identification susceptibility of genomic data privacy protection methods for shared data. Our results prove that the current set of privacy protection methods do not guarantee the protection of the identities of the data subjects. This work stresses that a new direction in the research and development of anonymity protection methods for genomic data must be undertaken. The next generation of privacy protection methods must take into account both social and computational interactions that occur in a complex data sharing environment. In addition, privacy protection methods need to provide proofs about what protections can and can not be afforded to genomic data.

Acknowledgements

The author wishes to thank various members of the Data Privacy Laboratory for their support and thought provoking discussions. This work was supported in part by the Data Privacy Laboratory at Carnegie Mellon University.

References

- [1] Altman RB and Klein TE. Challenges for biomedical informatics and pharmacogenomics. *Annu Rev Pharmacol Toxicol.* 2002; 42: 113-133.
- [2] Andrew SE, et. al. The relationship between trinucleotide (cag) repeat length and clinical features of Huntington's disease. *Nature* 1993; 4: 398-403.
- [3] Brinkman RR, et. al., The likelihood of being affected with Huntington disease by a particular age, for a specific CAG size. *Am J Hum Genet* 1997; 60: 1202-1210.
- [4] Burnett L, Barlow-Stewart K, Pros AL, Aizenberg H. The "GeneTrustee": a universal identification system that ensures privacy and confidentiality for human genetic databases. *Journal of Law and Medicine* 2003 May; 10(4): 506-513.
- [5] Churches T. A proposed architecture and method of operation for improving the protection of privacy and confidentiality in disease registers. *BMC Med Res Methodol* 2003 Jan 6; 3(1): 1.
- [6] Clayton EW. Ethical, legal, and social implications of genomic medicine. *New England Journal of Medicine* 2003 Aug 7; 349(6): 562-569.
- [7] NHS Executive. *Data Protection Act 1998 - protection and use of patient information.* Leeds: NHS Executive, 2000. (HSC 2000/009).

- [8] Department of Health and Human Services. 45 CFR (Code of Federal Regulations), Parts 160 – 164. Standards for privacy of individually identifiable health information, Final Rule. Federal Register: 67 (157): 53182-53273, Aug 12 2002.
- [9] Equifax-Harris. Health information privacy survey. Louis Harris and Associates, New York, NY, 1993.
- [10] Gaudet D, Arsnauld S, Belanger C, Hudson T, Perron P, Bernard M, and Hamet P. Procedure to protect confidentiality of familial data in community genetics and genomics research. Clin Genet 1999; 55: 259-264.
- [11] Gulcher JR, Kristjansson K, Gudbjartsson H, and Stefansson K. Protection of privacy by third-party encryption in genetic research. Eur J Hum Genetics 2000; 8: 739-742.
- [12] Gulcher JR, Kong A, and Stefansson K. The genealogic approach to human genetics. The Cancer Journal 2001 Jan/Feb; 7(1): 61-68.
- [13] Hall MA and Rich SS. Patients' fear of genetic discrimination by health insurers: the impact of legal protections. Genet Med 2000; 2: 214-221.
- [14] Malin B and Sweeney L. Determining the Identifiability of DNA Database Entries. In Proceedings of the Proc AMIA Symposium, 2000; pp 537-541.
- [15] Malin B and Sweeney L. Re-identification of DNA through an automated linkage process. In Proceedings of the AMIA Annual Fall Symposium, 2001; pp. 423-427.
- [16] Malin B and Sweeney L. Inferring genotype from clinical phenotype through a knowledge-based algorithm. In Proc of the Pacific Symposium on Biocomputing, 2002. pp 41-52.

- [17] Malin B, Sweeney L, Newton E. Trail re-identification: learning who you are from you have been. LIDAP-WP12. Carnegie Mellon University, Data Privacy Laboratory, Pittsburgh, PA. March 2003.
- [18] de Moor GJ, Claerhout B, de Meyer F. Privacy enhancing technologies: the key to secure communication and management of clinical and genomic data. *Meth Info Med* 2003; 42: 148-153.
- [19] National Association of Health Data Organizations, *NAHDO Inventory of State-wide Hospital Discharge Data Activities* (Falls Church: National Association of Health Data Organizations, May 2000).
- [20] Rothstein MA, ed. *Genetic Secrets: Protecting Privacy and Confidentiality in the Genetic Era*. New Haven: Yale University Press, 1997.
- [21] Sweeney L. Uniqueness of simple demographics in the U.S. population. LIDAP-WP4. Laboratory for International Data Privacy, Carnegie Mellon University. 2000.
- [22] Vaszar LT, Cho MK, Raffin TA. Privacy issues in personalized medicine. *Pharmacogenomics* 2003; 4(2): 107-112.
- [23] Wylie JE and Mineau GP. Biomedical databases: protecting privacy and promoting research. *Trends in Biotechnology* 2003 Mar; 21(3): 113-116.

Appendix A

Solving the family structure attack proof:

$$s_n = \sum_{i=0}^n (n-i+1) * (i+1)$$

$$s_n = \sum_{i=0}^n (ni + n - i^2 - i + i + 1)$$

$$s_n = n \sum_{i=0}^n i + n \sum_{i=0}^n 1 - \sum_{i=0}^n i^2 + \sum_{i=0}^n 1$$

Let $A = \sum_{i=0}^n i$, $B = \sum_{i=0}^n 1$, and $C = \sum_{i=0}^n i^2$. Then our formula is:

$$s_n = nA + nB - C + B$$

It should be obvious that $\sum_{i=0}^n 1 = n+1$, and thus:

$$s_n = nA + (n+1)^2 - C$$

A and C, we solve through induction. Solving A is simple.

$$\sum_{i=0}^n i = \frac{n(n+1)}{2}$$

Solving C may be less intuitive, but we'll see that $\sum_{i=0}^n i^2 = \frac{n(n+1)(2n+1)}{6}$. First, observe that this

equation holds in the base case when $n=0$. It should equal 1:

$$\sum_{i=0}^1 i^2 = 1$$

Now, let's show that it holds inductively:

$$\sum_{i=0}^{n+1} i^2 = \sum_{i=0}^n i^2 + (n+1)^2$$

$$\sum_{i=0}^{n+1} i^2 = \frac{n(n+1)(2n+1)}{6} + \frac{6(n+1)(n+1)}{6}$$

$$\sum_{i=0}^{n+1} i^2 = \frac{2n^3 + 3n^2 + n}{6} + \frac{6n^2 + 12n + 6}{6}$$

$$\sum_{i=0}^{n+1} i^2 = \frac{2n^3 + 9n^2 + 13n + 6}{6}$$

$$\sum_{i=0}^{n+1} i^2 = \frac{(n+1)(n+2)(2n+3)}{6}$$

Thus, plugging in our solutions to A and C :

$$s_n = n \frac{(n)(n+1)}{2} + n(n+1) - \frac{(n)(n+1)(2n+1)}{6} + (n+1)$$

$$s_n = \frac{3n^3 + 3n^2}{6} + (n+1)^2 - \frac{2n^3 + 3n^2 + n}{6}$$

$$s_n = \frac{3n^3 + 3n^2}{6} + \frac{6n^2 + 12n + 6}{6} - \frac{2n^3 + 3n^2 + n}{6}$$

$$s_n = \frac{n^3 + 6n^2 + 11n + 6}{6}$$