Brown University

Department of Computer Science &
Center for Computational Molecular Biology (CCMB)

Approaches in Genomic Privacy

Arun Das

Undergraduate Senior Honors Thesis
May 2018

Advisor: Professor Sorin Istrail

Submitted as part of the requirements for an Sc.B. in Computational Biology (Computer Science track) with Honors.

This thesis is dedicated to the four people who made this possible.

My incredible parents, Amit and Shobha Das, without whom I would never have gotten this far.

My advisor, Professor Sorin Istrail, for inspiring me to study Computational Biology.

And Helen Denisenko, whose support cannot be put into words.

Abstract

Genomic data contains extremely sensitive information about an individual, and its compromise or disclosure can have disastrous implications. However, in spite of these obvious risks, the current security practices employed to secure such data often fall woefully short.

This thesis aims to survey the current state of genomic privacy. It covers some of the challenges facing genomic privacy, including current threats and demonstrated attacks. It presents and reviews potential technical, cryptographic and policy solutions that may help guarantee the privacy and security of genomic data.

Contents

Chapter 1

Introduction

We live in the age of big data, with exabytes of it being generated and consumed every day by billions across the world. And yet, despite its abundance, we often neglect to secure data properly. While much data can be insignificant, and its security not vital, there are many types of data that must be secured in order to protect the rights and safety of the individuals they relate to.

One such form of data is genomic data. By its very nature, genomic data contains extremely sensitive information about an individual - it can hold the key to a person's predisposition to diseases, their likely response to different treatments, and even aspects of their identity, such as their ethnicity and relatives. Compromise of these data can subject their owner to social embarrassment, discrimination, or even targeted threats to their physical security. Worse still, the leakage of data on one individual simultaneously exposes data on all their kin with whom they shares genetic similarities [10]. Furthermore, in addition to the immediate risks to individuals from the accidental or malicious disclosure of their genomic data, such loss of privacy undermines public confidence and the willingness to share genomic data for research.

Current security practices, such as anonymization, often fall woefully short. For example, it has been shown that it is possible to uniquely identify an individual within a genome wide association study[9, 2], completely circumventing current security measures in place. The problem is made even harder by the fact that a genome itself is a unique identifier - it takes just 80 single nucleotide polymorphisms (SNPs) to uniquely identify one individual[13].

Therefore, it is not sufficient to simply remove personal identifiers - we must further restrict access to unencrypted (plaintext) genomic data. Such restriction on plaintext access must be balanced against the need to conduct search and other computations on genomic data (for research, medical or forensic purposes). Therefore, if we encrypt such data, we are restricted in our choice of encryption schemes, as the schemes used must support search and computation on encrypted data. Currently, the most promising approaches in genomic privacy use a combination of aggregated or obfuscated release techniques, such as differential privacy[5, 17, 7] or summary statistics, and privacy-enhancing cryptographic techniques, including homomorphic encryption, secure multiparty computation [11], and functional encryption [1]. However, some of these approaches have large computational or storage overheads, forcing us to make tradeoffs between security and efficiency.

This thesis aims to survey the current state of genomic privacy. It will review the work of leading researchers in genomic privacy, with a focus on security vulnerabilities identified by them and their suggested solutions. It will evaluate and compare existing cryptographic methods employed in genomic privacy, and consider how cryptographic techniques developed and used for other privacy-preserving applications could be applied to genomic data. Finally, it will include projections for the future of genomic privacy, in terms of both technologies and their likelihood of adoption.

Chapter 2

Background

The advent of cheaper genome sequence technology has allowed for tremendous progress in genomics. Since the sequencing of the first full human genome, a $300 million dollar venture at the turn of the millenium, the cost of genome sequencing has plummeted faster than even Moore's Law would have predicted. Today, the cost of sequencing a full human genome is a hundred thousand times less, with sequencing companies able to sequence an individual's genome for around $1000 (see Figure 2.1).

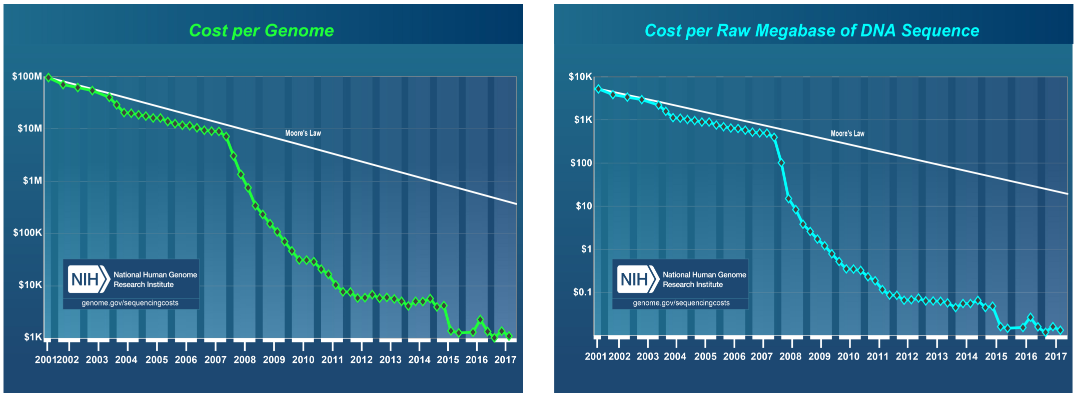


Figure 2.1: Cost of Genome Sequencing (*Source: NIH National Human Genomic Research Institute*)

The rapidly declining cost of genome sequencing has given rise to the age of genomic data. The past decade has seen nations, organizations and even individuals embark on ambitious initiatives to sequence large numbers of genomes, in an effort to better map, study and understand heredity and health. Two examples of this are the All Of Us Research Program[1], which aims to collect health and genetic data from one million American citizens, and Genomics England[2], a government project launched to sequence the genomes of 100,000 patients in British hospitals who are suffering from rare diseases or cancer.

Such initiatives have not only allowed us to collect vast amounts of genomic data, but also furthered the technologies we use to collect such data, resulting in more, higher quality sequence information. These vast banks of genomic information are primarily used for research purposes,

[1] https://allofus.nih.gov

[2] https://www.genomicsengland.co.uk

with the main goal often being advances in healthcare. Large biorepositories are used to draw insights into and further our understanding of the features and patterns of the human genome. The arrival of these large genomic datasets has also prompted the need for efficient and effective data sharing techniques. Partially driven by this need, the Global Alliance for Genomics and Health (GA4GH) was set up in 2013 to streamline the sharing of such data between organizations through the development of tools such as the Beacon Project[3] (which allows researchers to search if a certain allele exists in a dataset) and the Matchmaker Exchange[4] (for rare disease discovery).

In addition to government initiatives, we have also seen a flurry of activity in genomics in the private sector. These companies specialize in sequencing machines (e.g. Illumina or Oxford Nanopore), the storage, processing and analysis of genomic data (e.g. Google Genomics) and direct-to-consumer genetic testing (e.g. 23andMe and AncestryDNA)[3].

However, there are new dangers brought about by the rise of genomics. The widespread availability of such data exposes us to previously unseen ethical, security and privacy threats, many of which we have failed to address. The difficulty of anonymizing data renders many traditional sanitization techniques obsolete. Genomic data contains a plethora of information beyond just the sequence - an individual's data reveals much about their traits, heritage and disease predispositions. Furthermore, the risk of exposure of such data does not affect just the individual it was collected from - it contains significant information on the relatives, ancestors and descendants of the individual. As a result, the lifetime of genomic data is unlike almost any other - its lifespan is that of the individual who provided it, and to some extent, the lifetimes of all their relatives and descendants.

While many threats to genomic privacy remain theoretical, some attacks have been carried out in practice. One such attack was demonstrated by Homer et al. [9], who showed that it was possible to determine whether a specific individual was part of a case study group, simply by comparing the target's genome (or in many cases, a fragment of their genome) to the aggregate statistics of a reference population (available publicly) and that of the case study. The implications of such an attack were so drastic that it caused the Wellcome Trust and the NIH, who supplied the dataset used by Homer et al., to withdraw the dataset entirely.

Homer's attack was just the beginning. Since that attack, the scale of the genomic data we collect and store has grown exponentially, exposing us to new potential threats and attacks. In the next section, we will describe the most significant attacks on genomic privacy.

To counter these threats, we have to secure and protect genomic data better. A trivial solution would be to take all publicly available genomic datasets offline, thus restricting their access to authorized users only. However, such an action would stifle global collaboration and greatly hamper genomics research. Therefore, we must focus on other strategies that can grant us some measure of security while maintaining access. In Chapter 3, we outline some of the core goals of genomic privacy, and some of the requirements that guide the techniques and mechanisms we employ.

The most basic of these strategies use anonymization and encryption, thus removing or hiding identifiers that exist in genomic datasets. Such techniques are largely effective, but remain prone

[3] https://beacon-network.org/

[4] http://www.matchmakerexchange.org

to inference attacks, such as the ones demonstrated by Homer et al. [9] and Gymrek et. al [8]. We will study these attacks in greater detail in Chapter 4, and look at other attack mechanisms that have been used on genomic datasets.

Another technical solution would be to employ obfuscated or aggregated release techniques, such as the use of summary statistics or differential privacy. Such techniques would still allow for genomic datasets to be made available to the public, but would limit the precision of the insights that can be gained from the data. We will explore these techniques in Chapter 5.

The final technical solution would be to modify traditional cryptographic techniques to work with genomic datasets. Such techniques build secure frameworks for semi-trusted parties to interact with and share encrypted data. In Chapter 5, we will briefly cover three of these techniques - homomorphic encryption, multi-party computation, and functional encryption - and address their suitability for genomic data.

We must also consider non-technical solutions. This will largely consist of strengthening and expanding current legislation, regulation and best practices regarding the collection, storage and distribution of genomic data. Non-technical solutions will be briefly discussed towards the end of Chapter 5. In addition, we must strive to improve, through education, the public awareness of the risks facing genomic data, and emphasize the importance of future work in genomic privacy.

Finally, we also consider the future development of genomic privacy. In Chapter 6, we will include projections for the future of genomic privacy, balancing the need for such privacy with the public good arising from the sharing of genomic data.

Chapter 3

Desiderata of Genomic Privacy

As we search for a solution to the problems facing genomic privacy, it is important to define what properties a viable solution should have. Here, I outline some of the desired properties for any system that is designed for genomic privacy.

1. Many of the proposed technical approaches are thought to be viable simply because they work in polynomial time and/or space (in the size of the input). However, this does not take into account the scale of genomic data. Genomic datasets often contain thousands of individuals' data, and each datum can be millions of letters long - being polynomial in the size of this input can make a scheme still too slow or computationally expensive to be practical. Therefore, we must constrain feasible solutions not just in time and space complexity, but also bound the runtimes and memory/space requirements, to create practical and scalable tools.

2. As we will see in Chapter 5, many of the proposed schemes are very restrictive in what interactions they allow between researchers and the data they work with. Most only allow for a handful of queries to be performed, place restrictions on types or amount of data that may be accessed, and place limits on the types of results that may be returned. While such restrictions are essential in order to provide security guarantees, they are impractical for genomics research. In practice, the workflow of a genomics researcher is rarely so restricted - thoroughly investigating and understanding a genomic dataset often involves a wide range of queries on various subsets of the data, and requires flexible combination of different queries. Therefore, when developing mechanisms for genomic privacy, we should aim to place as few restrictions as possible on the ways in which scientists interact with genomic data.

3. Genomic data has a far longer lifespan than most forms of data - its immediate sensitivity is bounded by the lifetime of the individual it is collected from, and more loosely, by the lifetimes of their relatives and descendants. Genomic data is also truly immutable, meaning it must be compromised just once for it to be forever revealed (unlike passwords or security data that can be reset upon revelation). As a result, the schemes used to protect such data must match its longevity, and must be designed to withstand a lifetime of attacks.

4. In the case of genomic data, full disclosure is not the only risk. We must also secure it against partial disclosure, since even small sections of an individual's genomic information can contain very sensitive information. Therefore, the schemes we develop and deploy must totally secure genomic data, and prevent any form of disclosure.

5. It is tempting to rely completely on cryptographic techniques, and the theoretical security guarantees they provide, to protect genomic data. However, there are inherent dangers in

doing so. Cryptography is a flawed art, and systems are secure only for as long as a bug is not found in them - the systems that we consider to be secure are only really *secure to the best of our knowledge.*

Historically, cryptosystems that were thought to have been provably secure have rarely stood the test of time. Therefore, we cannot blindly place our trust in “proven” cryptographic systems - such systems are only secure against current adversaries, and given the current state of knowledge. They make no claim about their security against cryptographic advances or adversaries in the future, and it is not inconceivable to see many of these secure systems being broken in the near future[1].

Therefore, we must employ a combination of cryptographic and non-cryptographic techniques when securing our data, so that the systems we design do not fail as spectacularly when there are advances in cryptanalysis.

6. Finally, we must review how we approach consent in genomic privacy. Currently, a very opaque, static approach to consent is employed - participants consent to their data being included in a genomic dataset, and short of a data breach that may affect their data, they are never contacted again. In many cases, participants are poorly informed of the risks of participating in such scientific studies, the privacy-preserving mechanisms (if any) that are in place, or the availability of their data.

   Such systems must be replaced with more transparent, dynamic consent systems with increased participant involvement, and more granular researcher-participant interactions. Participants should have a say in which aspects of their data are shared, how much privacy they wish to have, how long their data may be used for, and whether they wish to continue to participate in the study or not.

[1] For example, if it is shown that $P = NP$, the vast majority of modern cryptography will become vulnerable.

Chapter 4

Past Attacks on Genomic Data

Brief Overview of Attack Mechanisms

We begin with an overview of the general types of attacks on genomic data. Attacks on genomic privacy tend to fall into three main categories[1]:

- **Identity Tracing**: Identity tracing attacks attempt to uniquely identify an anonymous DNA sample using quasi-identifiers from the dataset. Success is measured as the information obtained by the adversary relative to the size of the population the data is drawn from.

  Identity tracing attacks largely take the following forms [6]:

  Exploiting meta-data : Genomic datasets are often published with additional metadata (demographic details, criteria to participate in the study, additional health information, etc.), which can be exploited to trace the identity of an unknown genome in the sample. Demographic metadata is an especially potent source of identifying information; estimates suggest that the combination of date of birth, sex, and 5-digit zip code uniquely identifies more than 60% of US individuals[18][2].

  Genealogical triangulation : The development of impressive online platforms to search for genetic matches, made for a worldwide community who wish to understand their genealogy, has made genealogical triangulation a viable attack for identity tracers. An example of this attack is the surname-inference attack demonstrated by Gymrek et al. [8], where the attackers exploit the Y chromosome–surname correlation.

  Phenotypic Prediction : A little more far-fetched and a little less practical, it has been envisioned that the prediction of phenotypes from genetic data could be used as quasi-identifiers for tracing[14].

  Side-channel Leaks : Such attacks exploit unintentionally coded quasi-identifiers in datasets, instead of targeting the actual data that is made public. These attacks exploit factors such as filenames, numbering, hash values, and other basic computer security vulnerabilities, to discover further information about participants in a genomic dataset[3].

[1] In this thesis, we will focus on techniques that use data mining or the combination of distinct resources to learn private genomic information, and not basic computer security vulnerabilities.

[2] Using this approach, Sweeney [18] successfully identified the medical condition of William Weld, former governor of Massachusetts, using only his demographic data (date of birth, gender, and 5-digit ZIP code) appearing in the hospital records and in voter registration forms that are available to everyone.

[3] Sweeney et al. [19] discovered that the uncompressed files from the Personal Genome Project (PGP) have filenames that contain the actual name of the study participant.

- **Attribute Disclosure Attacks via DNA (ADAD)**: ADAD attacks harness genetic markers or characteristics to identify individuals and disclose further information about them. Some techniques include:

  n=1 scenario : The most basic scenario is one where the sensitive attribute in the dataset is associated with the genotype data of the individual. In this case, the adversary can simply match the genotype data that is associated with the identity of the individual and the genotype data that is associated with the attribute [6]. Such an attack requires a small number of SNPs to perform with high accuracy, thus making GWAS highly vulnerable.

  Summary Statistics : Such an attack might work as follows. Consider an extremely rare variation in the subject's genome - a non-zero allele frequency of this variation in a small study increases the likelihood that the target was part of the study, whereas a zero allele frequency strongly reduces this likelihood. An example of this exact attack was demonstrated by Homer et al. [9].

  Gene Expression : The crux of such an attack is to learn loci in gene expression profiles that are the most probable markers of a given genotype. Once such markers are learned, they can be used to compare anonymized gene expression datasets (such as the NIH's Gene Expression Omnibus (GEO)) to medical data with patient information.

- **Completion Attacks**: Completion of genetic information from partial data (genomic imputation) is a common problem, and, using a combination of linkage disequilibrium between markers and reference panels with complete genetic information, can be used to recreate missing genotypic values. This same approach can be used to "fill in the blanks" in sanitized datasets, where only partial DNA information is made available.

  Completion attacks use imputation to learn more information about targets and their relatives. This can be done through further analysis of the available genetic information (and harnessing linkage disequilibrium vis-a-vis reference panels to patch the gaps), extrapolating existing genomic information to infer the genotypes of relatives (who may or may not appear in the partially sanitized dataset), or using the available genomic information of relatives to infer information about an individual (thus completing the missing data).

Homer, 2008: Resolving Target Membership in a Complex Mixture

The first major attack on genomic data was demonstrated by Homer et al. in 2008 [9]. Resolving whether an individual's genomic DNA is present in trace amounts within a complex mixture (one containing DNA from many individuals) is of interest in many fields. However, before Homer's paper, identifying individuals who contribute less than 10% of a mixture was difficult and inaccurate, with the best available techniques relying on Short Tandem Repeats (STRs) or Mitochondrial DNA (mtDNA). The approach presented by Homer et al. is able to identify trace amounts ($< 1\%$) of DNA from an individual contributor within a complex mixture.

The attack uses the hundreds of thousands of SNPs on a high-density microarray (Affymetrix or Illumina) as a means to resolve trace contributions of DNA to a complex mixture. This technique can then be used to determine whether a specific individual was part of a case study group, simply by comparing the target's genome to the aggregate statistics of a reference population and of the case study. To do this, the allele probe intensities of the target individual are measured (using ratio of intensity measures from common biallelic SNPs), and compared to the intensities of the same alleles in the case study and the reference population. The idea here is that the unique combination of SNPs possessed by the target will be more similar to one of these groups, thus allowing the experimenters to make a prediction about the target's presence in or absence from in the case study.

The setup is as follows. The attacker knows of a case study group, and a reference population [4] (the control group), and is attempting to determine if a target individual (whose partial or whole genome is known) is present in the case study group. The attacker then computes the population allele frequencies for both the case study group and the reference population, as well as the allele frequencies (intensities) for the target individual. The distance between the individual and the two populations is then computed for all alleles (see Figure 4.1).

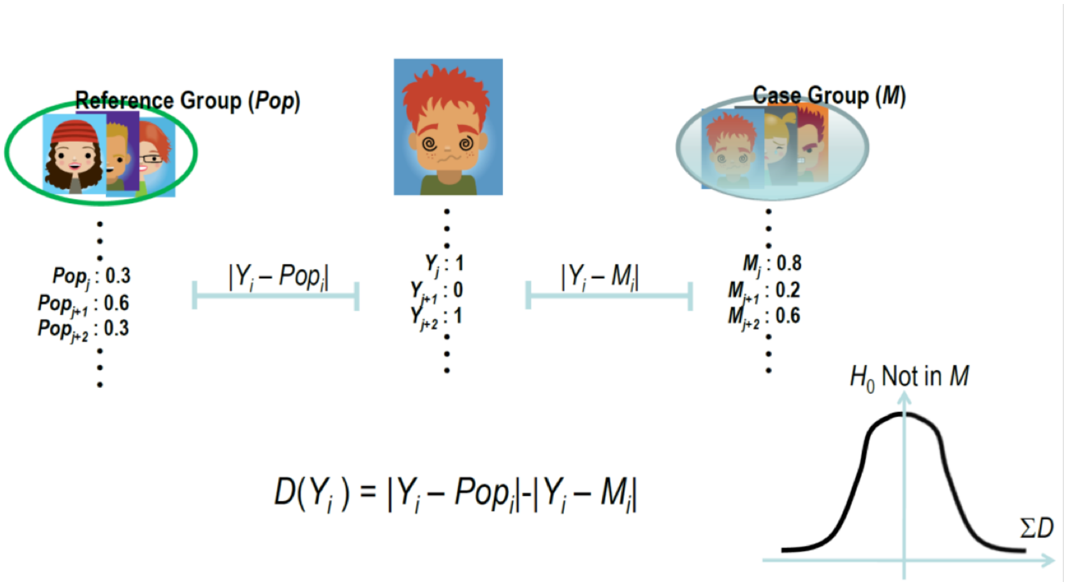
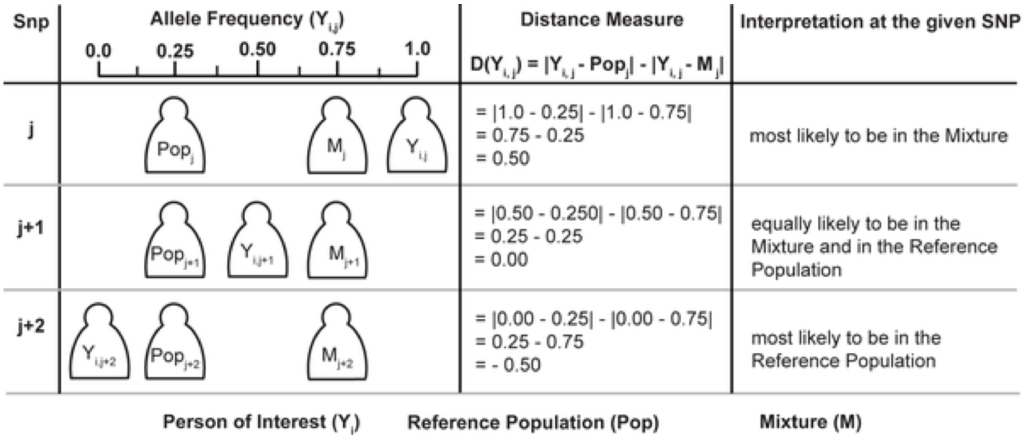


Figure 4.1: Homer's attack on a single target individual. Y_i is the allele intensity at the i^{th} allele in the target, Pop_i and M_i are the intensities of the same allele in the reference population and case study group respectively, and $D(Y_i) = |Y_i - Pop_i| - |Y_i - M_i|$ is the distance measure used. *(Source: Erman Ayday and Jean-Pierre Hubaux, ACM CCS 2016.)*

The distances across all alleles is summed, and the attacker can then make a prediction regarding the target's membership in the populations. If the distance is positive, the individual is predicted to be in the case study group, while a negative distance suggests that the individual is closer to the reference population (and therefore not in the case study group). A distance value close to 0 implies that the individual is equally likely to be in either population (see Figure 4.2).

In Homer's experiment, the mixtures used (as the case study groups) were composed by randomly sampling N individuals from the 58C Wellcome Trust Case-Control Consortium dataset of 1423 individuals (see [9] for the full list of mixtures).

[4] It is assumed that the reference population is accurately matched to the case study and the target.



Snp	Allele Frequency ($Y_{i,j}$)	Distance Measure $D(Y_{i,j}) = \lvert Y_{i,j} - Pop_j\rvert - \lvert Y_{i,j} - M_j\rvert$	Interpretation at the given SNP
j	Pop_j M_j $Y_{i,j}$	= |1.0 - 0.25| - |1.0 - 0.75| = 0.75 - 0.25 = 0.50	most likely to be in the Mixture
j+1	Pop_{j+1} $Y_{i,j+1}$ M_{j+1}	= |0.50 - 0.250| - |0.50 - 0.75| = 0.25 - 0.25 = 0.00	equally likely to be in the Mixture and in the Reference Population
j+2	$Y_{i,j+2}$ Pop_{j+2} M_{j+2}	= |0.00 - 0.25| - |0.00 - 0.75| = 0.25 - 0.75 = - 0.50	most likely to be in the Reference Population

Figure 4.2: Homer's attack on a single target individual: Y_{ij} is the allele intensity at the i^{th} allele in the j^{th} target, Pop_i and M_i are the intensities of the same allele in the reference population and case study group respectively. *(Source: Homer et al. [9])*

The findings of Homer's results were largely consistent with prior expectations. It was found that more SNPs allowed for greater resolution, and that the mixture fraction is the defining aspect of the problem - smaller fractions require far more SNPs to identify correctly. However, what was startling was the accuracy of the attack. To resolve mixtures where the target was <1% of mixture (with p-value $< 10^{-6}$), 25,000 SNPs are needed, while resolving mixtures where the target contributes anywhere from 10% to 0.1% takes between 10,000 and 50,000 SNPs. On all the mixtures tested in the experiment, the approach was universally successful, with the exception of mixtures where the target had relatives present (but even in this case, the approach could be slightly modified to utilize the relatives to resolve the target individual).

The effectiveness of Homer's attack shocked the NIH and the Wellcome Trust into action - the datasets used by Homer et al. were withdrawn, and only re-released after being fully anonymized. Homer's attack had a grave impact on GWAS studies as a whole - the realization that individuals may now be identified in publicly available experimental datasets has continued to haunt such studies ever since.

The effectiveness of Homer's attack was further formalized and quantified in Visscher et. al [20]. Visscher used likelihood ratios and linear regression to determine an upper bound on the power of such an attack to identify an individual in a complex mixture.

Wheeler, 2008: The complete genome of an individual by massively parallel DNA sequencing

An interesting result was obtained by Wheeler et al. in 2008 [26]. Unlike the other papers presented in this section, this result is not strictly an attack - instead, it just shows the power of modern sequencing technology.

In this paper, Wheeler et al. report the DNA sequence of a diploid genome of a single individual, (James D. Watson), using a combination of massively parallel DNA sequencing and comparison to a reference genome. The former is not too relevant to the problem of genomic privacy, but the latter is quite startling - just by comparing to the reference genome, Wheeler et al. identified 3.3 million single nucleotide polymorphisms, 10,654 of which cause amino-acid substitution within the coding sequence, small-scale insertion and deletion polymorphisms, and copy number variation resulting in the large-scale gain and loss of chromosomal segments [26]. These revelations allowed researchers to obtain sensitive details about the target, which in turn prompted sections of the individual's genome to be removed from publicly accessible repositories. The combination of cheaper sequencing and the abundance of reference genomic information has made it possible to completely sequence an individual's genome faster, cheaper and more accurately than ever before.

While the result shown in this paper is not strictly an attack, it is not hard to see its potential to be exploited. Obtaining a small amount of DNA from a target, sequencing this DNA (using techniques such as massively parallel sequencing), and then comparing the results obtained to existing reference genomes or genomic datasets, can allow an attacker to fully reconstruct, and then exploit, a target's genetic information. Therefore, it is imperative that genetic and genomic information is thoroughly protected.

GWAS and Aggregate Data Attacks

In a series of papers [21, 23, 24], it was demonstrated that it is possible to attack genomic privacy using aggregated data and summary statistics.

In [21], Wang et al. attempt to show that the threat demonstrated by Homer et al. has been greatly understated. In this paper, it is shown that individuals can actually be identified from even a relatively small set of statistics, such as those usually published in GWAS papers. Two attacks are presented - the first extends Homer's attack with a much more powerful test statistic (based on SNP correlation), and determines the presence of an individual using the statistics related to a few hundred SNPs. The second has the potential to lead to the total disclosure of the SNPs of hundreds of individuals participating in a study, just by studying information derived from the published statistics. Both these attacks are shown to be effective even with low precision statistics and with partial data, on both simulated and actual data.

Next, in [23], Wang et al. show that mining GWAS statistics threatens the privacy of a much wider population. To do this, Wang et al. provide a method to construct a two-layered Bayesian network whose goal is to reveal conditional dependencies between SNPs and traits in public GWAS catalogues. Once this network is developed, efficient algorithms are presented for two attacks - identity inference and trait inference. The targets of such an attack are not limited to participants of the GWAS study. These approaches are tested and evaluated, and the results show that it is possible to exploit unprotected GWAS statistics in order to identify individuals or derive previously hidden information.

Finally, in [24], Wang et al. demonstrate that even with differentially private GWAS statistics, there is still a risk for leaking individual privacy. Once again, a Bayesian network is constructed through mining public GWAS statistics, and trait and identity inference attacks are performed on both simulated and real human genetic data from 1000 Genome Project. These attacks attempt to infringe the privacy of not only GWAS participants, but also other individuals. The success of these attacks demonstrates that unexpected privacy breaches could occur and attackers can derive identity and private information from aggregated GWAS data.

Gymrek, 2013: Surname inference in public anonymized genomic datasets

A more recent attack was demonstrated by Gymrek et al. [8]. In the paper, it is shown that surname inference is possible from public anonymized genomic datasets using short tandem repeats (STRs) on the Y chromosome.

This attack is made possible thanks to a quirk in human societal norms. Surnames are paternally inherited in many human societies, resulting in co-segregation with Y-chromosome haplotypes. This relation has been taken advantage of by genealogy companies, who link patrilineal relatives by simply genotyping a handful of short tandem repeats[5] (STRs) on the Y chromosome.

Now, the existence of numerous genomic databases and surname projects across the globe, which together provide thousands of surname-haplotype pairs, has allowed this linkage to be exploited by attackers seeking to de-anonymize data. The attack itself is not new - in the past this linkage has been used to link the children of anonymous sperm donors to their biological fathers[8]. This linkage, combined with demographic information, can completely identify an individual.

In Gymrek et al., end-to-end identification, from a personal genome dataset to a individual or a set of individuals, is demonstrated. This identification process takes advantage of recreational genetic genealogy databases (YSearch, SMGF), as well as simple internet searches. These databases allow users to enter Y-STR alleles and search for matching records, and the records themselves contain information such as location, pedigrees, and potential other spellings of the surname.

The initial attack (searching the databases, computing the confidence of each returned record, and selecting the most likely one) achieved a success rate of just 12% - which is still dangerous when genomic data is at stake - with 5% of surnames incorrectly guessed. However, combining this data with demographic information allows narrowing down of the set of possible sample sources to just a handful of individuals. In scenarios where the genomic data is available along with the target's year of birth and state of residency (information not protected by HIPAA), using online public record search engines and US census data can reduce the search size to less than a dozen individuals.

Such an attack proves that anonymized data, even if sanitized thoroughly and correctly, is not immune to inference attacks, when other databases containing related information are available.

[5] Short (2-5 bases) sequences of DNA that are repeated numerous times in a head-tail manner.

Chapter 5

Protecting Genomic Data

In this section, we will look at some of the mechanisms that are currently employed to protect genomic data - in particular, we will consider the technical, cryptographic and policy mechanisms that are in place to secure genomic data. We will also consider some possible privacy-preservation techniques that could be applied or modified to secure genomic data.

Before considering the first mechanism, it is important to note that a trivial solution to most of the problems plaguing genomic privacy would be to place limits on the data that is made publicly available. For example, Sankararaman et al. [15] demonstrated how to limit the statistical power of Homer's attack by restricting the amount of data published. However, such a solution runs counter to the central goal of sharing genomic data - to further research into genomics. Therefore, we will only consider mechanisms that do not place limits on the amount of genomic data that is shared.

Similarly, it does not suffice to take a simple "security-by-obscurity" approach. Under this approach, the mitigation strategy is to simply remove obvious identifiers before making the dataset public. However, such schemes are based on the low probability of an adverse event, and therefore remain vulnerable in the face of a determined and motivated attacker. Furthermore, such a view is somewhat short-sighted - it is impossible to estimate future risks of adverse events[1].

The failings of security by obscurity are best described in Shannon's maxim (which is a reformulation of Kerchoff's principle), which states that an adversary knows the system they are attacking, and that "one ought to design systems under the assumption that the enemy will immediately gain full familiarity with them[2]". Therefore, it is important to design schemes that do not rely on the small chances of a breach, especially when knowledge of the workings of the scheme greatly amplifies these odds. That being said, there is nothing wrong with schemes utilizing secrecy - they just should not rely on it alone.

Access Control

The implementation of access control schemes could greatly improve genomic privacy [6]. Consider the following implementation, used by the NCBI's database of Genotypes and Phenotypes (dbGaP):

- Sensitive genomic data would be stored in a secure location, and all requests to access it would be screened and logged.

[1] "The Black Swan" - Nicholas Taleb

[2] Claude Shannon, 1985

- If the request is approved, the data may be accessed or downloaded, with the condition that the data will be viewed/stored under secure conditions, and no attempts will be made to identify individuals. Violating these conditions would result in access being revoked, as well as other potential penalties.
- Approved users are required to file frequent reports about the usage of the data, and report any adverse events.

While this system seems well-intentioned, there is a lack of oversight once the user has gained access to the data, and the reporting of breaches is left entirely up to the (potential malicious) user. Furthermore, revoking access and any other penalties may prevent the violating user from repeating their actions in the future, but does nothing to stop a breach from happening in the first place.

Alternatively, we could employ a trust-but-verify scheme, where users cannot download the data without restrictions, but can perform only certain queries depending on their access privileges. Such a system supports better early detection - malicious queries, actions or other anomalous behaviors can be flagged as they happen, and can be handled directly.

Yet another model that could be employed is let the participants of a study manage access control [6]. In such a model, participants grant access to their genomic data, bypassing the need for a data access committee. Such an implementation requires on-going communication between researchers and participants, and allows participants to modify their preferences whenever they desire. This implementation allows for higher levels of participant involvement, and makes the process of sharing genomic information more transparent. An example of this model is already in use by PEER (Platform for Engaging Everyone Responsibly)[3].

Summary Statistics

A widely-used obfuscated release technique is the use of summary statistics. Several studies ([7, 12]) have explored the differentially private release of common GWAS data summary statistics, such as allele frequencies, χ^2-statistics, p-values[7], or the location of variants[12].

Fienberg et al. propose new methods to release aggregate GWAS data without compromising an individual's privacy [7]. These methods are evaluated on both simulated data and a GWAS of canine hair length. However, it is found that a large amount of noise must be added even for the release of GWAS statistics from a small number of SNPs, meaning that these schemes are often impractical. These schemes also fall short on bigger and sparse data, where a summary statistic does not properly capture the complexity of the dataset. To counter these short-comings, the authors propose a differentially private algorithm for a specific form of penalized logistic regression. In this algorithm, noise is added to the analysis itself. However, this scheme is very much in its infancy - see [7] for more details.

[3] http://www.geneticalliance.org/programs/biotrust/peer

Privacy Metrics: k-anonymity, l-diversity, t-closeness

Another alternative is to better quantify the privacy offered by current anonymization techniques.

The first of these measures is k-anonymity, which is a property of anonymized data. A dataset is said to be k-anonymous if the information for any given person contained in the dataset cannot be distinguished from the information of at least $k-1$ other individuals. In other words, there exist at least k individuals with the same combination of attributes in the dataset, for all combinations of attributes. There are two common methods for achieving k-anonymity.

Suppression : Certain values of attributes are masked (e.g. replaced by a "*"). This may be all or some of the values that appear in a column of the dataset.

Generalization : Individual attribute values are replaced by broader categories. For example, the value "23" in the X column may be replaced by the range "$20 \leq X < 30$".

Meyerson and Williams (2004) [4] proved that optimal k-anonymity is an NP-hard problem, but heuristic methods (such as k-Optimize) and practical approximation algorithms (with an approximation guarantee of $O(\log k)$ do exist. However, since k-anonymity does not include any randomization, it is still susceptible to inference attacks. It is also not well suited for high-dimensional datasets, where a single individual ($k = 1$) can be unique identified very easily using the combination of attributes. Finally, use of k-anonymity can skew the results of a data set if it disproportionately suppresses and generalizes data points.

An alternative measure is l-diversity. l-diversity is a form of group-based anonymization that preserves privacy by reducing granularity. An equivalence class is said to have l-diversity if there are at least l "well-represented[5]" values for the sensitive attribute, and a table is said to have l-diversity if every equivalence class within it has l-diversity[6]. A practical example of l-diversity appears in location-tracking - a dataset containing individuals' location tracking data is said to be l-diverse if using all of the data in the dataset only allows a particular individual's true location to be narrowed down to one of l potential locations.

l-diversity was motivated by k-anonymity's susceptibility to inference attacks, and was designed to maintain all the privacy offered by k-anonymity while additionally maintaining the diversity of sensitive fields. As such, the techniques used to achieve l-diversity are similar to those used for k-anonymity.

Finally, we can consider t-closeness. t-closeness is an even more refined version of l-diversity, where the distribution of values taken on by an attribute are considered before modifying the attribute. This change was made to deal with the difficulty of protecting l-diversity against attribute disclosure (as not each value may display the same sensitivity, and rare positive attributes may give away more information than a common negative one).

The original paper that introduced t-closeness defines it as follows: "an equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said

[4] https://dl.acm.org/citation.cfm?id=1055591

[5] Relatively commonly occurring.

[6] From "t-Closeness: Privacy beyond k-anonymity and l-diversity (2007)"

to have t-closeness if all equivalence classes have t-closeness".

Whilst these three measures have not been explicitly developed for genomic data, they could be useful in helping quantify the privacy offered by current anonymization or obfuscating schemes.

Cryptographic Techniques

In this thesis, we review three cryptographic techniques, and focus on their potential for use in genomic privacy.

The techniques we review are:

- Differential Privacy (DP): DP is a technique that aims to maximize the accuracy of queries from statistical databases while minimizing the chances of identification.
- Secure Multi-Party Computation (MPC): MPC techniques allow for parties to jointly compute a function over their inputs while keeping those inputs private.
- Homomorphic Encryption (HE): HE is a form of encryption that allows for computation on ciphertexts, generating an encrypted result, that when decrypted, matches the result of the same operations performed on plaintext.

Differential Privacy

Work in differential privacy was motivated by Dalenius' 1977 desideratum for statistical databases, which states that nothing about an individual should be learnable from the database that cannot be learned without access to the database. This result was shown to be impossible [5] - it is indeed impossible to publish information from a private statistical database without revealing some amount of private information, and the entire database can be revealed using a very small number of queries. The main obstacle to this desiderata, and the cornerstone of the impossibility result, is the presence of auxiliary information that is available to the adversary from sources other than the statistical database. To demonstrate the power of such information, consider the following example [5].

Suppose the exact height of an individual is deemed to be highly sensitive and private information, and that the revelation of such information would constitute a privacy breach. Assume the existence of a database that contains the average heights of women of different nationalities. An adversary who has access to this database and the auxiliary information "Helen is four inches taller than the average American woman" can learn Helen's height. An adversary without access to the database learns nothing new about her height. This example works regardless of whether Helen is in the database or not, thus proving that Dalenius' goal (of providing privacy in statistical databases using semantic definitions) is impossible. A formal proof of this impossibility result is given in [5], but the crux of it is as follows - an adversary (modeled as a Turing Machine) with access to auxiliary information can bypass the privacy mechanisms in place on any statistical database with non-negligible probability, while an adversary without such information cannot.

This prompted the creation of Differential Privacy by Dwork in 2006 [5]. In this paper, Dwork presents differential privacy as a measure which, intuitively, captures the increased risk to one's privacy incurred by participating in a database, and provides a mechanism that can be used to achieve any desired level of privacy under this measure.

The formal definition of this measure is as follows - A randomized function $\mathcal{K}$ gives ϵ-differential privacy if for all data sets D_1 and D_2 differing on at most one element, and all $S \subseteq Range(\mathcal{K})$,

$$Pr[K(D_1) \in S] \leq exp(\epsilon) \times Pr[K(D_2) \in S]$$

Such a mechanism addresses concerns that participants in a study may have about the leakage of their information. The removal of their information from a study would make no outputs significantly more or less likely. This definition can be extended to groups of individuals - c individuals concerned about the leakage of their collective data can bound this probability above by at most $exp(\epsilon)$. The scheme is designed to disclose aggregate information for large groups, and therefore disintegrate with large c.

An alternative explanation of ϵ-differential privacy comes from the frequentist perspective [25]. In this perspective, $exp(\epsilon)$ can be considered as the bound on the power-to-significance ratio of any statistical test an adversary may use to determine the disease status of a participant based on ϵ-differentially private data [17].

The privacy mechanism used to achieve differential privacy is rather simple - it works by adding appropriately chosen random noise to the answer $a = f(x)$, where X is the database and f is the query function. The magnitude of the random noise added is chosen as a function of the largest change a single participant could cause in the output to the query function (referred to as the sensitivity of the function). More formally, for $f : \mathcal{D} \rightarrow R^d$, the sensitivity of f is

$$\delta f = \max_{D_1, D_2} ||f(D_1) - f(D_2)||$$

for all D_1, D_2 differing in at most one element [7]. The privacy mechanism $\mathcal{K}_f$, for a query function f, computes $f(X)$ and adds random noise by sampling from a scaled symmetric exponential distribution with variance σ^2, using the density function

$$Pr[\mathcal{K}_f) = a] \propto exp(-||f(X) - a||/\sigma)$$

. For $f : \mathcal{D} \rightarrow R^d$, $\mathcal{K}_f$ gives $(\frac{\delta f}{\sigma})$-differential privacy.

Wasserman & Zhou [25] show that one way to ensure the preservation of information in differential privacy is to require that the distance between the probability distribution functions - the true (based on the raw data) and the empirical (computed from the released data) - remain small. The rate at which this distance goes to zero (with increase in the sample size of the released data) is a measure of the accuracy of different privacy mechanisms. In particular, for differential privacy methods using the exponential mechanism to add noise, the accuracy is proportional to the rate at which the empirical distribution concentrates into a small ball around the true distribution. Unfortunately, most of the privacy mechanisms investigated in the paper converge slower than the optimal (minimax) rate. The statistical view of differential privacy is explained in more detail in [25].

The pioneering implementation of Differential Privacy in genomic privacy was demonstrated in Berger et al. [17]. In their paper, Berger et al. introduce a novel variant of differential privacy

[7] This scheme works best when δf is small

tailored to genomic databases, with the aim of protecting private phenotype information (such as disease status) while correcting for population stratification.

Berger et al. introduced the concept of Phenotypic Differential Privacy (PDP), which they define as "a formal definition of privacy that attempts to preserve private information about individuals (such as disease status)". Just like differential privacy itself, PDP requires the choice of a privacy parameter/budget ϵ, which controls the level of privacy guaranteed to all participants in the study - the closer to zero it is, the more privacy is ensured, while the larger it is, the weaker the privacy guarantee is [17].

The framework introduced in Berger et al. performs GWAS while using principles of differential privacy to protect phenotype information. Privacy-preserving GWAS results, based on EIGENSTRAT and linear mixed model (LMM)-based statistics (both of which correct for population stratification), can be produced. The differentially private statistics (PrivSTRAT and PrivLMM), are tested on both simulated and real GWAS datasets, and it is found that meaningful results can be efficiently returned on two types of queries (GWAS statistics at SNPs of interest and highly associated SNPs of diseases of interest) without compromising privacy. Further details and results of this framework can be found in [17].

The development of such a framework is very encouraging. The main computational bottleneck in the framework comes from the computation of the original statistics. This suggests that such an implementation is well positioned to take advantage of the latest advances in GWAS analysis, which could give us more computationally efficient methods for generating the statistics. The use of differential privacy in genomic datasets would allow for widespread sharing and exchange of genomic information, and the quantified privacy guarantees it provides would hopefully encourage increased participation in genomic studies.

Another (earlier) implementation of the use of differential privacy on genomic datasets can be found in Johnson et al. [12]. In their 2013 paper, Johnson et al. present privacy-preserving exploratory data mining algorithms, where analysts do not need to know a priori which or how many SNPs to consider in GWAS datasets.

Johnson et al. develop such algorithms for the computation of the number and location of significantly associated SNPs (using a distance-based method, explained below), the significance of a statistical test between a disease and a given SNP, and measures and structures of correlations between SNPs [12]. The scheme presented in the paper supports the following GWAS queries:

1. the number of significant SNPs for a given p-value,
2. the location of the k SNPs with the highest p-value,
3. the location of the longest correlation block[8],
4. the p-value of a given SNP, and
5. the correlation between two SNPs.

Using a combination of such queries, an analyst is able to translate a reasonable idea of the number of significant SNPs into a quantified understanding of their identities, p-values and correlations

[8] See [12].

[12], thus enabling them to better select the statistical tests and correlation measures to use.

To guarantee differential privacy, Johnson et al. develop a distance-score mechanism, in order to better deal with the complicated output spaces of some queries and address their high sensitivity[9]. This distance-score mechanism is used instead of a Laplace-mechanism (for adding noise to queries), thus preserving the utility of the queries. The exponential mechanism (outlined above) is then used, and differential privacy follows. Finally, the algorithms presented are then tested on real-world datasets (in this case, a GWAS dataset on Irritable Bowel Syndrome), and are shown to guarantee differential privacy while returning reasonably accurate results[10].

However, schemes like [17] and [12] are few and far between, and implementations of differential privacy are very limited in the queries and statistics they support [11]. Furthermore, it is hard to find a balance between a reasonable ϵ-parameter and meaningful results. From a research perspective, a low privacy parameter is beneficial, as it allows for more noise-less, and therefore more accurate, inference - however, such a privacy parameter would expose participants to potential identification. Similarly, a high amount of privacy would protect participants of a study, but could render the results of queries to the database meaningless (if too much noise is added).

Based on the results of these studies, it is clear that there is a need for another mechanism that can both satisfy differential privacy and add less noise to released statistics, or the need for a different model to supplement differential privacy.

Multi-Party Computation (MPC)

The latest breakthrough in using MPC for genomic privacy has come from Bejerano et al. [11]. The approach in this paper is based on Yao's Protocol for Two-Party Computation [27].

Yao's protocol can be demonstrated using a simple example. Consider a scenario where two parties, Alice and Bob, each possess a secret number (a and b respectively) between 1 and 10. Alice and Bob wish to compute whether $a \geq b$, without revealing their secret number to each other. To do this, they construct 100 indistinguishable boxes, one corresponding to each possible (a, b) pair. Each of these boxes has two locks on it - one corresponding to one of the 10 keys Alice has (one for each value of a), and one for each of the 10 keys Bob has (one for each value of b). In each box, they leave a note. In 55 of these boxes, the note reads "Alice wins" (as in 55 cases, $a \geq b$), and in the remaining 45 boxes, "Bob wins".

Once the initial setup is completed, Alice and Bob leave the room housing the boxes. Alice then re-enters the room, and tries to unlock boxes until she unlocks the 10 boxes that correspond to the value of a she has chosen. Alice leaves the room, and Bob enters - he sees the the 10 partially unlocked boxes, and attempts to unlock them all, until he finds the one that corresponds to his choice of b. He can read the note inside the box, letting him know if $a \geq b$, and then leave the room. Alice re-enters, reads the note, and leaves the room.

[9] See figure 8.2 for the function, and details of the distance function can be found in [12].

[10] The authors note that the distance-score mechanism being used is only an approximation, and that computational advances here would help increase accuracy.

[11] Such schemes also often require very large numbers of participants to guarantee acceptable levels of privacy or utility.

This concludes the protocol. At the end of it, Alice and Bob both know the truth value of $a \geq b$, but neither has learned each other's secret. This is Yao's Protocol for Two-Party Computation.

Bejerano et al. attempt to leverage this protocol in the exchange of genomic data [11]. The paper introduces a proof-of-concept cryptographic implementation to allow two parties exchange genomic data. First, patient genomes are converted into simple-valued vectors to reveal the causative variants (this conversion process is described below). Yao's protocol is then used to perform the desired computation without revealing any participant's input [11]. To apply Yao's protocol, it is assumed each individual involved in a study has private access to her own exome/genome. If identification of a causal variant is the goal, each individual is given a variant vector of all possible missense and nonsense variants in the human genome (28,413,589 bases for the exome) - the individuals then note, using "true" or "false", whether they have each variant or not. If a causal gene is being identified, each individual is given a gene vector of 20,663 genes - the individuals then note "1" next to a gene if they have one or more rare functional variants in the gene, and "0" otherwise [11].

The framework defines three simple Boolean operations - `INTERSECTION, SETDIFF`, and `MAX` - that are used for patient diagnosis. From [11]: "`INTERSECTION` of two variant vectors reveals all rare functional variants that two parties share, `SETDIFF` of an affected and unaffected individual's variant vector allows us to discard variants seen in healthy individuals, ... `MAX` operation can be used to find a gene containing rare functional mutations in the greatest number of affected cases". The privacy guarantees provided by the system are quantified by the "protection quotient" - the fraction of private information that is exposed neither to the other participants nor to the entity running the computation. In the version of the protocol used in this paper, this is the ratio of the total number of patient variants withheld from the output to the total number of patient variants input into the computation[12] [11].

The framework proposed by Bejerano and et al. is tested by using the three secure operations over actual patients with causal Mendelian variants. The results show that the protection quotient of these operations is at least 97.1%, with the `MAX` operator achieving a measure of 99.7% - a full summary of the results can be seen in [11], and in the appendix of this paper. The performance of this scheme is rather impressive - even on single-threaded execution, the performance is very good, and comparable to the best techniques currently employed in genomic privacy.

While this paper demonstrates that MPC is a potential avenue for achieving genomic privacy, there are still numerous limitations. The current protocol allows only for two-parties to exchange information - scaling this to n-parties is currently unfeasible. Furthermore, even with two-parties, this framework is highly restrictive - it only supports the three given operations. While these operations may suffice for the questions being investigated in some studies, they are unlikely to cover all the research questions that are asked when studying genomic data. At this time, it is primarily a proof-of-concept work, and shows that the genomic privacy may have a future in multi-party computation.

Homomorphic Encryption (HE)

An HE scheme enables arbitrary computation on encrypted genomic data and allows for the reuse of the same encrypted input across multiple computations. It allows for large encrypted datasets

[12] Unprotected diagnoses have a protection quotient of 0%.

to be uploaded and stored on an untrusted cloud, and allows for computations to be performed on it without access to decryption keys. A good initial foray into the use of this technique for bioinformatics is by Dowlin et al. [4], who present Microsoft Research's "Simple Encrypted Arithmetic Library"[13], a set of tools to be used for experimentation and research purposes.

Implementations of basic homomorphic encryption (e.g. systems that only allow one operation) have existed for years. A homomorphic encryption solution which allows an unlimited number of pairs of operations (such as addition and multiplication) can be used to compute any circuit, and is therefore referred to as fully homomorphic (such a scheme was first presented by Gentry in 2009). Numerous schemes have been proposed, and for practical applications, only homomorphic encryption schemes which allow for a fixed amount of computation are used. Knowing this computation in advance leads to improved parameters, and better computational and storage efficiency.

Fully Homomorphic Encryption (FHE) provides a versatile solution for many privacy problems and scenarios. A single data owner can encrypt and store data securely in an untrusted cloud, and the use of either private or public key versions can allow many parties to upload new data to this repository, or perform computations on encrypted data. The scope of the access by other parties is fully determined by the data owner.

However, as tempting as it is to see FHE as a potential silver bullet for the problems facing genomic privacy, it is important to recognize its limitations. Existing FHE implementations are quite inefficient, and do not scale well when used on large genomic datasets. For reference, the Yao's Protocol used by Bejerano et al. is 5000-10000 times faster than the latest FHE scheme [11]. Furthermore, many of the proposed schemes remain as proof-of-concept works, without practical and usable implementations. Therefore, FHE is not currently a viable option for genomic privacy - the design of new implementations, or the optimization of existing ones, is needed before this cryptographic technique can see widespread use in genomic privacy.

Other Cryptographic Techniques

Another major technique that may in the future be employed for genomic privacy is Functional Encryption (FE). FE is a form of public-key encryption where possessing a secret key allows one to learn a function of what the ciphertext is encrypting. As of 2012, schemes existed that support arbitrary functions, but no schemes built specially for genomic data have been developed.

Finally, no effort in privacy in the modern age is complete without at least a brief mention of blockchain. Currently, blockchain has only been suggested as a means for individuals to monetize access to their genomic data[14]. The crux of the current implementations is as follows - individuals commit their data to the blockchain, maintained by the company offering this service, thus creating a (supposedly) immutable entry of their data. External parties then pay for access to this history, and some portion of this money is paid to the individual who owns the genetic data. The potential benefits of such a scheme are obvious. The use of a blockchain would allow parties to reach distributed consensus without the need for a middleman. This would allow for direct interaction between the providers of the data (participants) and the consumers (researchers), and would move power and control towards participants. However, such schemes still have their issues. The first

[13] http://sealcrypto.codeplex.com/

[14] For an example of such a scheme, see Nebula Genomics.

is their potential inefficiency - the amount of computing power needed to start and maintain the distributed system is still high. There may also be significant opposition to such a scheme, as it reduces the influence and control governments and institutions have on individuals' data. These blockchain based genomic privacy approaches are still very much in their infancy, and are not yet ready for research data-sharing. Therefore, this paper will not delve further into blockchain-based schemes.

Ethics and Policy Changes

Since its inception, genomic research has struggled to find the balance between genetic privacy and data access. As explained by Shi [16], "technological advances are followed and accompanied by concerns, debates, and controversies on a wide range of topics in ethics, regulations, and laws regarding protection and preservation of genetic privacy". While such laws and regulations often lag behind, their impact on personal genomics should not be understated.

Protecting genomic anonymity is extremely difficult, thanks to the ability to combine patient data (protected or not) with auxiliary information. As a result, it is more important than ever to focus on educational efforts to train relevant parties (students, researchers, medical practitioners) about the tools available for genetic privacy protection. Large human genome projects such as the Human Genome Project, HapMap Project and the 1000 Genomes Project have already started providing education on the ethics of genomics. There is also a need to improve education and awareness on the technical aspects of genomic privacy - many of the technical and cryptographic solutions currently used and proposed (including those in this thesis) run the risk of being too inaccessible for widespread use. To deal with this situation, we must strive to develop more easily usable and accessible tools, and create documentation and resources to help learn and use these tools.

As far as regulation goes, one milestone development in genomic data regulation in the United States was the Standards for Privacy of Individually Identifiable Health Information[15] in the Health Insurance Portability and Accountability Act of 1996 (HIPAA). This standard addressed the use and disclosure of an individual's health information by organizations subject to this regulation[16], and helped provide standards for individual privacy rights to understand and control the use of their health information[16].

However, HIPAA has its limits. Health and medical information not originating from covered organizations are not covered by HIPAA, meaning that commercial sequencing and genetic screening companies generate large amounts of sensitive and identifiable data that is not regulated by HIPAA. This is also true for other consumer-generated health information, such as fitness trackers, mobile apps, and social media - the data from these sources are not afforded the same regulation. Furthermore, the standard only protects identifiable health information, with no restrictions on the use or disclosure of de-identified data. This is potentially problematic, as the ability to combine publicly accessible auxiliary information (including metadata such as age, race or location) with de-identified information can be used to re-create and disclose sensitive information (such as in the attack by Gymrek et al.).

[15] i.e. the Privacy Rule

[16] Health providers, insurers, data clearing houses and their business partners.

These shortcomings have prompted proposals to revise HIPAA to handle recent advances - in particular, there is a need to regulate data generated by entities not covered by the original standard. There have also been calls to revamp consent forms, with studies showing that informed consent in genomic studies is often non well-informed - participants are rarely aware of the true risks of data exposure, or educated about the protection techniques (or lack thereof) that will be employed to protect their data. There is also the need to employ more dynamic and transparent consent techniques that allow participants to update their preferences or revoke their consent, and be more involved in the entire data-sharing process.

There is also a need to regulate private data exchange in cross-institutional studies [22]. Data may be shared between research institutions, sequencing facilities, and insurance and health care providers, and the lack of a regulatory framework exposes this sensitive data to numerous risks. Inconsistent protection and privacy-preserving measures can greatly increase the risk of data exposure. It is entirely possible that one party deems an aspect or attribute of the data to be not worthy of protection, while other parties consider it to be sensitive - in such a scenario, any privacy-preserving techniques employed by the first party are undone by the revelation of this information by latter parties. This lack of a unified front is greatly beneficial to adversaries, who can exploit inconsistencies to bypass the safeguards and measures in place to protect genomic data.

The prospects for regulation are further exacerbated when these institutions are spread across the globe, and the national laws governing the disclosure and use of genomic data in each of these institutions may be largely incompatible. A potential solution would be the creation of international regulations governing the use and disclosure of genomic data. However, such regulation remains extremely unlikely - vast differences in medical, health care and research systems, privacy and data-protection laws, and global politics all stand in the way of such regulation being crafted and signed into action.

Finally, there is a need for more advocacy for genomic privacy. The past decade has seen the meteoric rise in the prominence and standing of digital rights and privacy groups, which in turn has fostered somewhat of a privacy revolution. The actions of such groups and individuals has furthered public awareness and the relevant parties to address these concerns, and in some cases, even draft legislation that deals with the biggest threats to privacy. Groups such as the Electronic Frontier Foundation (EFF) have recently expanded into genomic privacy[17], and other groups are showing signs of doing so too, but there is still a long way to go.

[17] https://www.eff.org/issues/genetic-information-privacy

Chapter 6

Conclusion

The emergence of cheap genomic sequencing technologies has resulted in vast amounts of genomic data becoming publicly available. This, combined with recent advances in computer science, has ushered in the genomics age - we now have extremely efficient algorithms to analyze vast genomic datasets in search of insights that will further biology, healthcare, and other related fields. However, in the search for these insights, it is tempting to ignore the sensitivity of the genomic data we are collecting and sharing. Genomic privacy seeks to protect these data, and the privacy of the individuals they are collected from.

This thesis surveyed the current state of genomic privacy, from the threats and attacks facing it, to the current security practices used to protect genomic data, to potential technical, cryptographic and policy solutions that could be employed to safeguard this data.

Unfortunately, while the existing mechanisms to provide genomic privacy are clearly valuable, none of the proposed solutions are perfect. Simpler measures such as anonymization or access control provide surface level security, but do not hold up against even a mildly persistent adversary. The cryptographic approaches, such as multiparty computation or homomorphic encryption, remain our most secure options, but they are too limited in the actions and queries they support, and computationally too expensive, for widespread use. Obfuscated release techniques such as summary statistics or differential privacy greatly limit the information that an adversary can obtain from the datasets, but they also limit the insights researchers can gain from interacting with the data. Even the non-technical solutions, such as reworking the consent systems currently in place or drafting new legislature and regulations to protect genomic data, are a long way from being implemented.

Finding the balance between privacy and utility is difficult - the collection, use and distribution of genomic data has brought about unparalleled progress, and it is often hard to justify why privacy concerns (many of which are still theoretical) should be placed ahead of potential progress. Even if such a balance can successfully be struck, the ever changing nature of genomics research, coupled with the endless progress in computation and cryptography, would render any solution short-lived; it would either fail in the presence of a previously unknown adversary, or survive long enough to become outdated or poorly suited to the newest goals in research[1]. Therefore, it seems unlikely that we see a complete solution to genomic privacy (i.e. one that satisfies the desiderata presented in Chapter 3) in the very near future.

[1] I am reminded of a quote from Batman - "You either die a hero, or live long enough to see yourself become the villain".

However, this should not diminish the importance of research on genomic privacy. The dangers to genomic data are very real, and the longer they go unchecked, the more we risk the privacy and safety of participants of genomic studies and datasets. It is vital that we recognize the sensitivity of this data, and that we consider the privacy of the people who contribute it to be of the utmost importance. The schemes and suggestions proposed in this paper, while flawed in many regards, still represent an improvement over the current approaches being employed to safeguard genomic data. These techniques should be treated and employed as makeshift solutions, and used until improvements are made to them, or better approaches are developed. In the near term, a side-by-side comparison of different genomic privacy techniques - DP, MPC and HE - in terms of their computational efficiency and privacy preservation (measured using k-anonymity and other privacy metrics) on standard datasets appears to be a worthwhile direction for research.

There are many promising signs. Techniques such as multi-party computation and homomorphic encryption, are well-poised to take advantage of the latest breakthroughs in computer science, which could render them computationally viable for widespread use. There may be ways to fine-tune obfuscated release techniques such as summary statistics or differential privacy so that we achieve reasonable and well-quantified tradeoffs between privacy and utility. Simpler approaches, such access control and anonymization, which may be vulnerable when used individually, can be combined to create systems that offer greater security guarantees[2]. Increased education and public awareness of the sensitivity of genomic data has caused a push for more legislation governing the collection, use and disclosure of genomic information, and for a revamp of traditional consent systems. Finding workable solutions for the problems facing genomic privacy remains an achievable goal that we must strive to work towards, and I look forward to seeing what we can accomplish over the coming years.

[2] This is the “defense in depth” approach to computer security.

Chapter 7

Bibliography

[1] Erman Ayday and Jean-Pierre Hubaux. “Privacy and Security in the Genomic Era”. In: *ACM SIGSAC Conference on Computer and Communications Security* (2016), pp. 1863–1865. DOI: doi:10.1145/2976749.2976751.

[2] Rosemary Braun et al. “Needles in the Haystack: Identifying Individuals Present in Pooled Genomic Data”. In: *PLoS Genetics* 5.10 (2009). DOI: doi:10.1371/journal.pgen.1000668.

[3] Emiliano De Cristofaro. *SoK: A Critical Analysis of Genome Privacy Research.* University College of London.

[4] Nathan Dowlin et al. “Manual for Using Homomorphic Encryption for Bioinformatics”. In: *Proceedings of the IEEE* 105 (2017), pp. 552–567. DOI: doi:10.1109/JPROC.2016.2622218.

[5] Cynthia Dwork. “Differential Privacy”. In: *33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006).* Vol. 4052. Springer Verlag, 2006, pp. 1–12. ISBN: 3-540-35907-9.

[6] Yaniv Erlich and Arvind Narayanan. “Routes for breaching and protecting genetic privacy”. In: *Nature Reviews Genetics* 15 (2014), pp. 409–421. DOI: doi:10.1038/nrg3723.

[7] Stephen E. Fienberg, Aleksandra Slavkovic, and Carline Uhler. “Privacy Preserving GWAS Data Sharing”. In: *2011 IEEE 11th International Conference on Data Mining Workshops* (2011). DOI: doi:10.1109/ICDMW.2011.140.

[8] Melissa Gymrek et al. “Identifying Personal Genomes by Surname Inference”. In: *Science* 339 (2013), pp. 321–324. DOI: doi:10.1126/science.1229566.

[9] Nils Homer et al. “Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays”. In: *PLoS Genetics* 4.8 (2008). DOI: doi:10.1371/journal.pgen.1000167.

[10] Mathias Humbert et al. “Reconciling Utility with Privacy in Genomics”. In: *WPES ’14 Proceedings of the 13th Workshop on Privacy in the Electronic Society* (2014), pp. 11–20. DOI: doi:10.1145/2665943.2665945.

[11] Karthik A. Jagadeesh et al. “Deriving genomic diagnoses without revealing patient genomes”. In: *Science* 357.6352 (2017), pp. 692–695. DOI: doi:10.1126/science.aam9710.

[12] Aaron Johnson and Vitaly Shmatikov. “Privacy-preserving data exploration in genome-wide association studies”. In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’13.* ACM Press, 2013. DOI: 10.1145/2487575.2487687. URL: https://doi.org/10.1145/2487575.2487687.

[13] Zhen Lin, Art B. Owen, and Russ B. Altman. "Genomic Research and Human Subject Privacy". In: *Science* 305 (5681 2014), p. 183. DOI: `doi:10.1126/science.1095019`.

[14] Christoph Lippert et al. "Identification of individuals by trait prediction using whole-genome sequencing data". In: *Proceedings of the National Academy of Sciences* 114.38 (Sept. 2017), pp. 10166–10171. DOI: `10.1073/pnas.1711125114`. URL: `https://doi.org/10.1073/pnas.1711125114`.

[15] Sriram Sankararaman et al. "Genomic privacy and limits of individual detection in a pool". In: *Nature Genetics* 41.9 (Aug. 2009), pp. 965–967. DOI: `10.1038/ng.436`. URL: `https://doi.org/10.1038/ng.436`.

[16] Xinghua Shi and Xintao Wu. "An overview of human genetic privacy". In: *Annals of the New York Academy of Sciences* 1387 (2017), pp. 61–72. DOI: `doi:10.1111/nyas.13211`.

[17] Sean Simmons, Cenk Sahinalp, and Bonnie Berger. "Enabling Privacy-Preserving GWASs in Heterogeneous Human Populations". In: *Cell Systems* 3 (2016), pp. 54–61.

[18] Latanya Sweeney. "Simple Demographics Often Identify People Uniquely". In: 671 (Jan. 2000).

[19] Latanya Sweeney, Akua Abu, and Julia Winn. "Identifying Participants in the Personal Genome Project by Name". In: *Harvard University, Data Privacy Lab 1021* (2013).

[20] Peter M. Visscher and William G. Hill. "The Limits of Individual Identification from Sample Allele Frequencies: Theory and Statistical Analysis". In: *PLoS Genetics* 5.10 (2009). DOI: `doi:10.1371/journal.pgen.1000628`.

[21] Rui Wang et al. "Learning Your Identity and Disease from Research Papers: Information Leaks in Genome Wide Association Study". In: *Proceedings of the 16th ACM Conference on Computer and Communications Security*. CCS '09. Chicago, Illinois, USA: ACM, 2009, pp. 534–544. ISBN: 978-1-60558-894-0. DOI: `10.1145/1653662.1653726`. URL: `http://doi.acm.org/10.1145/1653662.1653726`.

[22] Shuang Wang et al. "Genome privacy: challenges, technical approaches to mitigate risk, and ethical considerations in the United States". In: *Annals of the New York Academy of Sciences* 1387.1 (Sept. 2016), pp. 73–83. DOI: `10.1111/nyas.13259`. URL: `https://doi.org/10.1111/nyas.13259`.

[23] Y. Wang, X. Wu, and X. Shi. "Using aggregate human genome data for individual identification". In: *2013 IEEE International Conference on Bioinformatics and Biomedicine*. 2013, pp. 410–415. DOI: `10.1109/BIBM.2013.6732527`.

[24] Yue Wang et al. "Infringement of Individual Privacy via Mining Differentially Private GWAS Statistics". In: *Big Data Computing and Communications*. Ed. by Yu Wang et al. Springer International Publishing, 2016, pp. 355–366. ISBN: 978-3-319-42553-5.

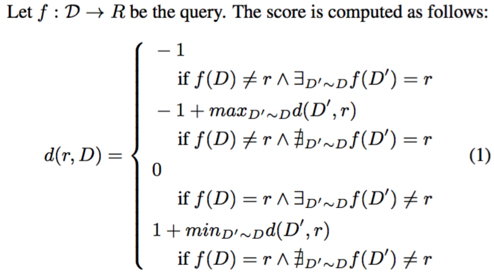
[25] Larry Wasserman and Shuheng Zhou. "A Statistical Framework for Differential Privacy". In: *Journal of the American Statistical Association* 105.489 (2010), pp. 375–389.

[26] David A. Wheeler et al. "The complete genome of an individual by massively parallel DNA sequencing". In: *Nature* 452 (2008), pp. 872–877. DOI: `doi:10.1038/nature06884`.

[27] Andrew C. Yao. "Protocols for secure computations". In: *23rd Annual Symposium on Foundations of Computer Science* 00 (1982), pp. 160–164. DOI: `doi:10.1109/SFCS.1982.88`.

Chapter 8

Appendix



Let $f : \mathcal{D} \to R$ be the query. The score is computed as follows:

$$d(r, D) = \begin{cases} -1 \\ \quad \text{if } f(D) \neq r \wedge \exists_{D' \sim D} f(D') = r \\ -1 + max_{D' \sim D} d(D', r) \\ \quad \text{if } f(D) \neq r \wedge \nexists_{D' \sim D} f(D') = r \\ 0 \\ \quad \text{if } f(D) = r \wedge \exists_{D' \sim D} f(D') \neq r \\ 1 + min_{D' \sim D} d(D', r) \\ \quad \text{if } f(D) = r \wedge \nexists_{D' \sim D} f(D') \neq r \end{cases} \qquad (1)$$

Figure 8.1: The distance-score measure used in Johnson et al. - see [12] for more details.