Quantifying Genomic Privacy via Inference Attack with High-Order SNV Correlations

Sahel Shariati Samani*, Zhicong Huang†, Erman Ayday‡,
Mark Elliot*, Jacques Fellay†, Jean-Pierre Hubaux†, Zoltán Kutalik§
*The University of Manchester, United Kingdom
† École Polytechnique Fédérale de Lausanne, Switzerland
‡ Bilkent University, Turkey
§ Lausanne University Hospital, Switzerland

***Abstract*—As genomic data becomes widely used, the problem of genomic data privacy becomes a hot interdisciplinary research topic among geneticists, bioinformaticians and security and privacy experts. Practical attacks have been identified on genomic data, and thus break the privacy expectations of individuals who contribute their genomic data to medical research, or simply share their data online. Frustrating as it is, the problem could become even worse. Existing genomic privacy breaches rely on low-order SNV (Single Nucleotide Variant) correlations. Our work shows that far more powerful attacks can be designed if high-order correlations are utilized. We corroborate this concern by making use of different SNV correlations based on various genomic data models and applying them to an inference attack on individuals' genotype data with hidden SNVs. We also show that low-order models behave very differently from real genomic data and therefore should not be relied upon for privacy-preserving solutions.**

I. Introduction

The rapid progress in genomic research and application raises serious privacy concerns. Various privacy problems have been revealed, and various privacy-preserving solutions have also been proposed. Many solutions focus on preventing an adversary from accessing individuals' sensitive genome sequences [1], [2], [3]. However, as described by Wang et al. [4], who show that an adversary can learn an individual's identity from published p-values (of statistics indicating the relation between SNVs and a disease) and linkage disequilibrium (correlation among SNVs) data, the outcome of genomic computation might leak sensitive genomic information. The privacy loss due to the availability of LD is also studied by some other work in genomic privacy [5], [6].

In genetic research, it is believed that higher-order correlations produce better models of genome sequences [7], [8], [9]. This implies that the above work can lead to more privacy loss if the higher-order correlations are integrated. Also, making use of the higher-order correlations, researchers can upgrade the existing privacy-preserving solutions that are originally designed on low-order correlations, and protect genomic data against stronger adversaries. For instance, Humbert et al. [10] propose a method to optimize the research utility of an individual's shared SNPs while satisfying privacy constraints from both an individual and his family members, by incorporating pairwise LD into part of their algorithm. So, by adapting the privacy constraints, we can, in principle, enhance the protection against a stronger adversary that uses high-order correlations for inference attacks. In this paper, considering different genomic data models, we demonstrate the privacy implications of different SNV correlations empirically with various inference results. Our work will help researchers to further understand the genomic privacy problems and design robust privacy-preserving solutions for genomic data even if information about high-order genomic correlation is available to an adversary.

The complex high-order correlation arises from the pattern of genetic recombination on chromosomes. Researchers have proposed and used methods to model genome sequences directly based on genetic recombination [11], enabling them to mimic the high-order correlations among SNVs. In this paper, we perform inference attacks with different genomic data models, including a first-order Markov model based on pairwise LD, multiple high-order Markov models built on real genomic datasets, and a genetic recombination model. The results show that high-order correlation indeed provides more information for inference attacks than low-order models.

The contribution of this paper can be summarized as follows:

- Based on several different genomic data models, we perform inference attacks to quantify individuals' genomic privacy;
- To further understand the characteristics of genomic data, we project the real genomic data and synthetic data generated with different models to two dimensions using principal component analysis, thereby visualizing and quantifying high-order correlation effects.

II. Technical Preliminaries

A. Genomic Background

In this section, we introduce the genomic background necessary to understand our methodology. Figure 1 provides an overview of the terminology.

This research was undertaken while Erman Ayday was at Ecole Polytechnique Fédérale de Lausanne.

DOI 10.1109/SPW.2015.21

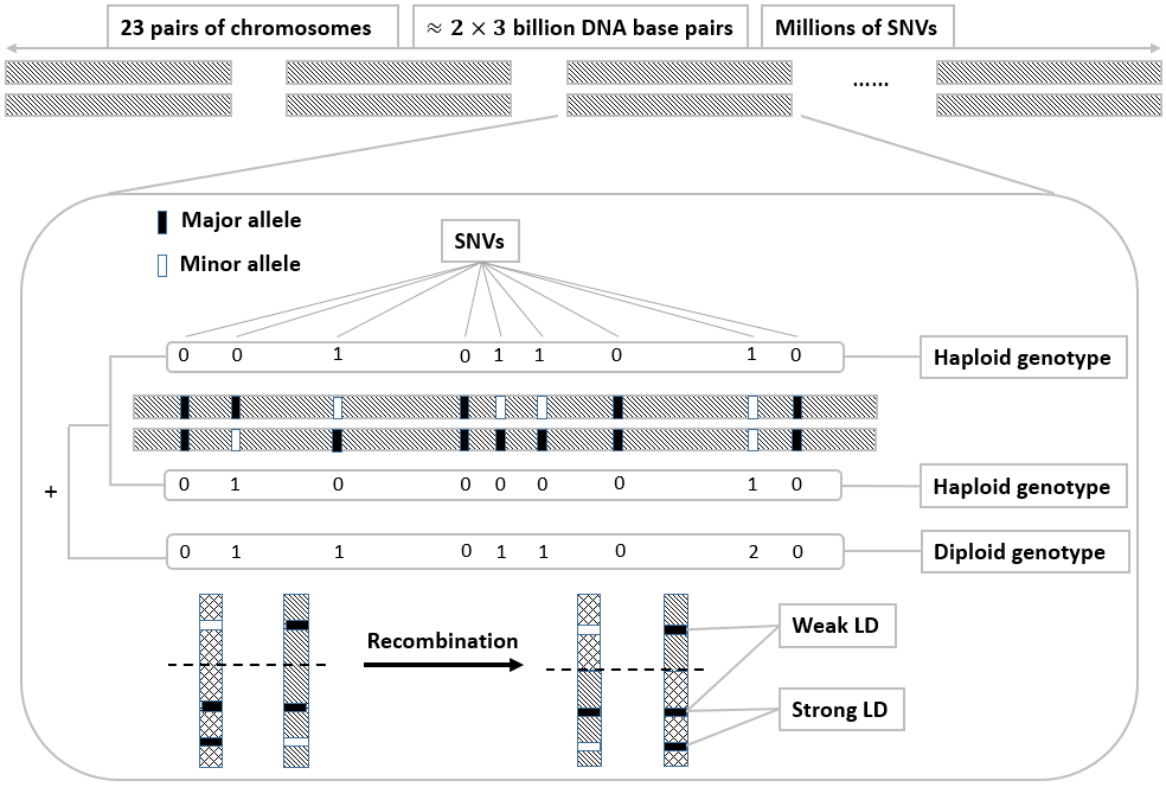


Fig. 1: Genomic background. Each of human's 23 pairs of chromosomes contains a large amount of genetic information, including millions of Single Nucleotide Variants (SNVs), the most common genetic variation. The alleles on one chromosome are collectively referred to as the haploid genotype, and the pairs of alleles on a pair of chromosomes are called the diploid genotype. When two SNVs located on the same chromosome are far from each other, they are likely to be separated by a recombination event during the meiotic process, which leads to a weak Linkage Disequilibrium (LD) between them.

1) Single Nucleotide Variant: Even if most of the sequence of any individual genome is identical to the reference human genome, each of us has about four millions differences, called variants. The most common genetic variants are single nucleotide variants (SNVs), where different alleles (A, T, C, or G) are observed at the same chromosomal position. Most SNVs are bi-allelic, i.e. there are only two possible alleles at that position: a major allele observed at higher frequency and a minor allele observed at lower frequency. For simplicity, our work only considers autosomal chromosomes, which are always inherited in pairs. As a consequence, there are three possible states for each SNV, i.e., (0; 1; 2) , depending on the number of minor alleles it carries.

2) Haploid and Diploid Genotypes: In this paper, we consider an individual's genomic data as a sequence of SNVs — called the diploid genotype — each of which takes values in $\{0, 1, 2\}$. As shown in Figure 1, an individual's haploid genotype is a sequence of alleles on one chromosome, in contrast to the diploid genotype on a pair of chromosomes.

3) Linkage Disequilibrium: Due to the inheritance mechanism, and in particular to the recombination process happening during meiosis, there is a non-random correlation between some alleles. Indeed, alleles that are close to each other on the same chromosome are more likely to be inherited together than would be expected by chance. This is known as linkage disequilibrium (LD). The correlation between pairs of SNVs has been well defined in several human populations, however the high-order correlation we use in this paper has been less explored.

4) Recombination: During meiosis, the pairs of parental chromosomes exchange DNA segments, leading to a novel combination of alleles that is passed on to progeny. The process is called recombination. The strength of the LD in a region depends on the recombination rate, which is variable across the genome. Genetic variants mapping to regions with a lower recombination rate are in stronger LD. Hence, genomic researchers propose modeling LD based on the recombination rates among SNVs, which we discuss in Section III-A3.

B. Adversary Model

In this paper, we consider a scenario where a victim contributes his genotype for research, or uploads it for medical test, or simply shares it for recreational purposes such as ancestry finding. Due to privacy concerns, he might want to hide some sensitive SNVs, such as those related to a genetic disease. An adversary is assumed to observe the victim's genotype with hidden SNVs. His objective is to infer the

missing SNVs. For such an inference attack, the adversary has access to necessary resources in order to build various genomic data models, such as allele frequencies, linkage disequilibrium, genetic recombination rates and sampled diploid and haploid genotypes in the same population as the victim. A well-known example in such a scenario comes from James Watson's decision to share his personal genome by releasing it on a publicly accessible scientific database except for his APOE gene (a gene associated with late onset Alzheimer's disease)[1]. However, Nyholt et al. show that Dr. Watson's APOE risk status can be easily and accurately predicted using publicly available data, such as HapMap data [12]. Another example is genomic data sharing websites where users publish some of their genetic variations online, such as OpenSNP [13]. Though not considered here, it is worth noting that a stronger adversary might also have access to pedigree and phenotypic information about the target. We leave exploration of this inference attack combining familial relationships, phenotypic data and high-order SNV correlations for future work.

C. k^{th}-Order Markov Chain

In a probabilistic model of genome sequences, each sequence can be represented as a sequence of genetic variants (SNVs) SNV_1, SNV_2,$\cdots$, which are ordered increasingly by their physical genomic positions. Each SNV_i takes a value from the set $\{0, 1, 2\}$ representing major homozygous, heterozygous, and minor homozygous diploid genotypes, respectively.

A k^{th}-order Markov chain is a sequence of random SNVs where the probability of SNV_i taking a particular value is only dependent on the values of k preceding SNVs.

$$P_k(\text{SNV}_i) = P(\text{SNV}_i|\text{SNV}_{i-1}, \text{SNV}_{i-2}, \cdots, \text{SNV}_{i-k}) \quad (1)$$

To use Markov chains to model genome sequences, we need to build a model which estimates the probability of SNV_i being equal to 0,1 or 2. For instance, in a 1^{st}-order model, the probability of each value in position i depends on the value of the previous position and therefore 9 probabilities need to be computed for each position: $P(0|0), P(0|1), \cdots, P(2|2)$. Note that in the genomics field, researchers are used to defining a hidden Markov chain that is useful for genotype imputation, and we will also introduce such a concrete model in Section III-A3. But be aware that it represents a totally different meaning from the Markov model we define above.

III. Methodology

In this section, we formalize our methodology by decomposing it into the several key components necessary for an inference attack to take place.

A. SNV Correlation Modeling

In genomics, researchers have proposed various methods to model the correlation across the genome sequence. Here we describe some state-of-the-art options.

[1]James Watson is one of the co-discoverers of the double-helix structure of DNA in 1953.

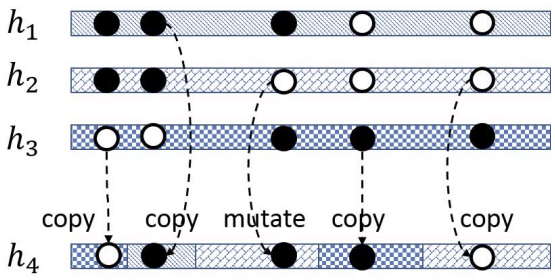


Fig. 2: An example (adapted from [11]) showing how the haploid genotype h_4 is interpreted as an imperfect mosaic of a given set of haploid genotypes $\{h_1, h_2, h_3\}$, based on recombination and mutation. Each column of circles represents a SNV locus on one chromosome, with colors black and white denoting two different alleles. The imperfect nature (mutation) of the copying process is exemplified at the third locus, whereas all other parts are exact copies from the existing haploid genotypes. Note that this shows just one possible process to get h_4 from $\{h_1, h_2, h_3\}$, and since there are many other possibilities, the purpose of this model is to compute the probability of observing h_4 by taking all the possible underlying processes into account, which constitutes a hidden Markov model.

1) Using Published Allele Frequencies and Linkage Disequilibria: For each SNV at position i, we can compute the probability of SNV_i being equal to 0,1 or 2 using published allele frequencies (AFs). Furthermore, LD relations are represented pairwise in literature and thus they can be used together with allele frequencies to compute the joint probability of SNV_i and SNV_j. Note that the joint probability of three SNVs (or more) is not computable with only pairwise LD. For each SNV_i, there could exist more than one LD relation; however in our method, to compute the probability of SNV_i being equal to 0,1 or 2, we consider only the previous adjacent SNV that usually has the strongest LD with SNV_i. In other words, in this method we build a 1^{st}-order Markov model using published AF and LD data.

2) Using Genotype Datasets: In general, higher-order Markov models should perform better than lower-order models and we would like to use the highest possible order. As LD relations are only provided pairwise, public AF and LD data do not capture higher-order correlations (if they exist) and it is not sufficient to model genome sequences using k^{th}-order Markov chain where $k > 1$. To build a higher-order Markov model, publicly available sequences can be used as training data in order to build k^{th}-order models for different values of k. Assume that we have N genome sequences as our training data. Let $F(\text{SNV}_{i,j})$ represent the frequency of subsequence $\text{SNV}_{i,j}$ that contains SNVs between SNV_i and SNV_j. The k^{th}-order model is then built by computing:

$$P_k(\text{SNV}_i) = \begin{cases} 0 & \text{if } F(\text{SNV}_{i-k,i-1}) = 0 \\ \frac{F(\text{SNV}_{i-k,i})}{F(\text{SNV}_{i-k,i-1})} & \text{if } F(\text{SNV}_{i-k,i-1}) > 0 \end{cases} \quad (2)$$

3) Using Genetic Recombination Rates: High-order SNV correlation is a result of different genetic recombination rates on different positions across the genome sequence. This provides a different method to model the high-order correlation, where researchers relate LD patterns to the underlying recombination rate [11]. Assume we have a set of t haploid genotypes $\{h_1, h_2, \cdots, h_t\}$. The model is developed to find the conditional distribution of the next observed haploid genotype, $Pr(h_{t+1}|h_1, \cdots, h_t)$. An example is shown in Figure 2. Each allele of h_{t+1} can be thought of as having been created by "copying" (an exact copy, or an imperfect one, leading to a mutation) the corresponding part of h_1, h_2, $\cdots$, or h_t. Intuitively, we think of h_{t+1} as having recent shared ancestry with the haploid genotype that it copied in each segment. The copying process is further assumed to be Markov along the chromosome. In other words, assuming one part of h_{t+1} comes from h_i, the next adjacent part could be copied from any of the t haploid genotypes and the jumping probability depends on the recombination rate between these two parts. Note that an extremely large recombination rate makes the two parts become independent, equivalent to a low LD value. To compute the probability of observing a particular haploid genotype h_{t+1}, we must sum over all possible event sequences of recombination and mutation that could lead to h_{t+1}. The Markov assumption allows us to do this efficiently, using standard forward-backward algorithm for hidden Markov models [14]. Note that in the forward algorithm, we could compute the probability of observing a prefix sequence of h_{t+1}, from the first position to the i^{th} position, denoted by $h_{t+1}^{1,i}$. Hence the conditional probability of observing the i^{th} allele, Allele_i, given all preceding alleles is computed as:

$$P(\text{Allele}_i|\text{Allele}_{i-1}, \cdots, \text{Allele}_1) = \frac{P(h_{t+1}^{1,i})}{P(h_{t+1}^{1,i-1})} \quad (3)$$

More details can be found in [11]. The model extension from haploid genotype to diploid genotype is available in [15] whose algorithm was implemented in a genotype imputation software called IMPUTE. In our work, we adapted the algorithm to our framework and implemented it in Python.

B. Inference Attack

We define the inference attack as finding the value of unknown SNV_i given the probabilistic modelling of genome sequences which represents the probability of SNV_i taking each value of the set $\{0, 1, 2\}$. In other words, we assume that the attacker has a genome sequence with some unknown SNVs and a probabilistic model of genome sequences. Then given the genome sequence and the model, the attacker estimates the value of unknown SNV_i. The overall framework is shown in Figure 3.

To model this attack, we first split the given dataset of genome sequences into the training and test dataset. We then use the training dataset to build different models using different methods. Next, in each sequence in the test dataset, we hide a specified number (denoted by s) of SNVs selected randomly and use each model to predict the hidden SNVs. The indices of the hidden SNVs are $x_1, x_2, \cdots, x_s$. For each hidden SNV, the predicted value is the one with the highest conditional probability in the corresponding model. In the end, we estimate the error in inferring the value of the hidden SNVs using different models. To this end, we quantify the average estimation error as follows:

$$E = \frac{\sum_{i=0}^{s} |\text{SNV}_{x_i}^p - \text{SNV}_{x_i}^r|}{s}, \quad (4)$$

where $\text{SNV}_{x_i}^p$ represents the predicted value of SNV_{x_i} and $\text{SNV}_{x_i}^r$ refers to the value of SNV_{x_i} in the real dataset.

IV. Evaluation

In this section, we first evaluate the performance of inference attack by considering the different methods to model genome sequences and then compare the average estimated errors. Then to better compare the characteristics of the different genomic data models, we randomly generate samples with each of the models and observe the resultant distributions. We reduce the dimensionality of the samples and project the real and the generated ones into a two-dimensional space for the purposes of visualization.

A. Dataset

The dataset used in these experiments is a publicly available one from the HapMap project [16]. It comprises the diploid genotypes for non-redundant SNP[2] assays of the chromosome 22 of 165 HapMap subjects, who are all Utah Residents with Northern and Western European Ancestry (CEU). We used the genotypes from phase III released in May 2010, including 17715 diploid genotypes in each sequence. A haploid genotype dataset, including 200 haploid genotypes that comes from the same population, was used to build the recombination model. Allele frequencies, pairwise linkage disequilibria, and recombination rates were also used to build the corresponding models.

B. Inference Attack Results

In this experiment, we first split the whole dataset, including 165 sequences, into the training and the test dataset. We randomly selected 100 sequences as the training data and used the other 65 sequences as the test data. To build our models, we used the training dataset and six methods, the 0^{th} through 4^{th}-order Markov chain models (M0, M1, M2, M3 and M4 respectively) and the recombination model (RM). In addition, we used allele frequencies and linkage disequilibrium data to model genome sequences of CEU population (M1-LD). Next, to perform the inference attack and estimate the inference error, we followed the following steps:

(i) hid 1771 randomly selected SNVs, which is about 10% of the total number of the SNVs, in each genome sequence in the test dataset.
(ii) used each model to predict the hidden SNVs.
(iii) measured the estimated error as described in Section III-B.
(iv) repeated steps 1 to 3 ten times.

[2]Single Nucleotide Polymorphism. It is a similar term as SNV, except that SNP usually refers to SNV with minor allele frequency larger than 0.01.

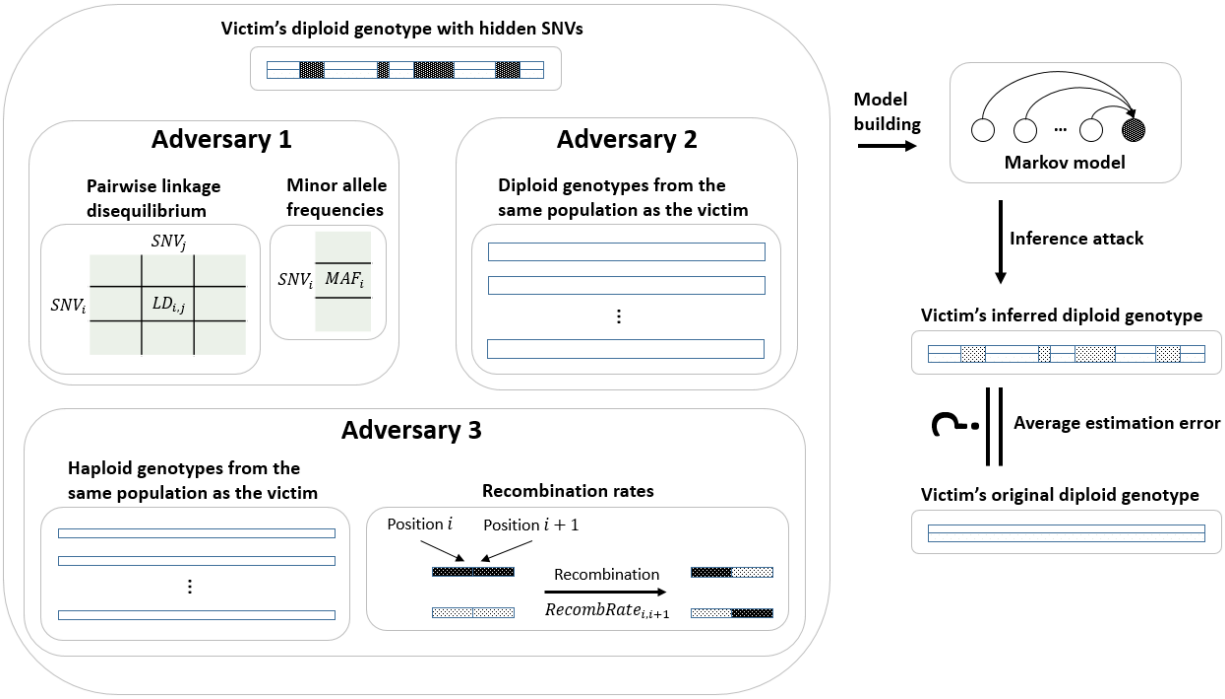


Fig. 3: Framework to quantify genomic privacy via inference attack with SNV correlation. To model the correlation, adversaries build the k^{th}-order Markov model based on various types of knowledge, such as allele frequencies, pairwise linkage disequilibrium, recombination rates, diploid genotype and haploid genotype datasets (e.g. the HapMap Project [16]). The average estimation error is used to evaluate the performance of different models.

(v) averaged the estimated errors for each genome sequence.

Figure 4a illustrates the average estimation error of inferring hidden SNVs based on different models on the test data.

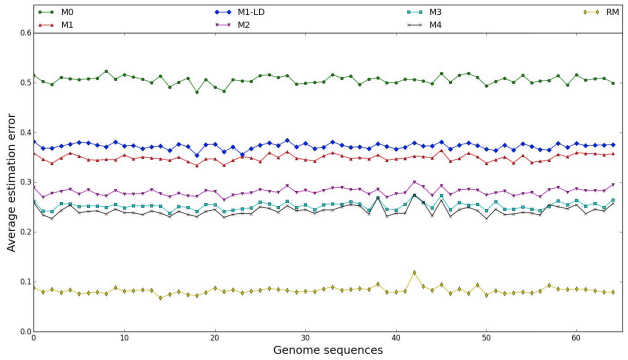
As Figure 4a indicates, by increasing k, the average estimation error decreases, which illustrates that the higher-order Markov chain could increase the accuracy of the attack. However, for $k > 3$, not much improvement can be observed, which shows the limitation of building k^{th}-order Markov model on diploid genotype datasets. Indeed, the plots for $k = 5$, 6 and 7 were heavily overlapping with that for $k = 4$ and they were not included in the figure to improve the visibility of the existing curves. Furthermore, it can be seen that the recombination model performs much better in predicting hidden SNVs as it considers all the correlations between SNVs to generate the model. Moreover, Figure 4a illustrates that the Markov models with $k > 1$ built using the training dataset and the recombination model perform much better than the model built using AF and LD data and they can predict hidden SNVs more accurately.

We also performed the same experiments but hid a larger number of SNVs (7086, nearly 40% of the total number) and the result can be seen in Figure 4b. As expected, the overall performance of different models are slightly worse when a larger number of SNVs is required to be inferred. Yet, the average estimation errors are mostly below 0.1 with the recombination model. The inference results with recombination model convey an important message: individuals' genomic data can be accurately inferred even when a large part of data is hidden. It also indicates a promising improvement over state-of-the-art genomic data inference attacks that only consider pairwise SNV correlations, such as the attack proposed by Humbert et al. in kin genomic privacy, taking familial relationships into account [6]. We leave this to future work.

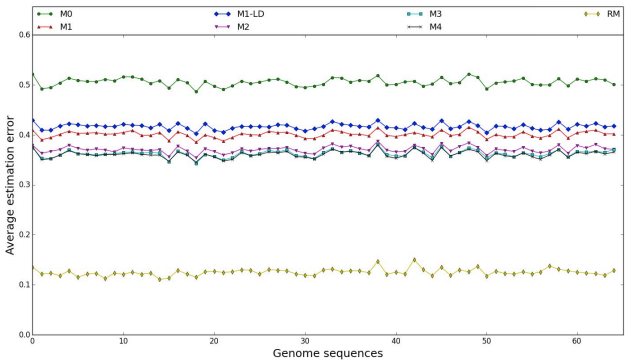
As seen in the above results, the recombination model provides a relatively more accurate estimation of the hidden SNVs. Figure 5 shows the overall performance of this model in a comprehensive way. We hid different percentages of SNVs and computed the corresponding estimation errors after using the recombination model. With only SNV correlations, the result indicates the power of an adversary by making use of a good genomic data model.

C. Visualizing and Quantifying High-Order Correlation Effects

To further understand the characteristics of genome sequences and the effects of high-order correlations between SNVs, we conducted experiments to observe the distribution of sample data generated with different models. In the first experiment, we generated 100 random samples using 1^{st}-order Markov model, the model based on public AFs and LDs, and the genetic recombination model. We then applied principle component analysis (PCA) to the real data (165 samples) in order to reduce the dimensionality of the data for better visualization. We extracted the first two principal



(a) Average estimation error on 10% hidden SNVs



(b) Average estimation error on 40% hidden SNVs

Fig. 4: Average estimation error using different models to infer (a) 10% unknown SNVs, and (b) 40% unknown SNVs. M0, M1, M2, M3 and M4 represents 0^{th}, 1^{st}, 2^{nd}, 3^{rd}, and 4^{th}-order Markov chain built on the diploid genotype dataset; M1-LD represents 1^{st}-order Markov chain built with public pairwise LD; RM represents recombination model.

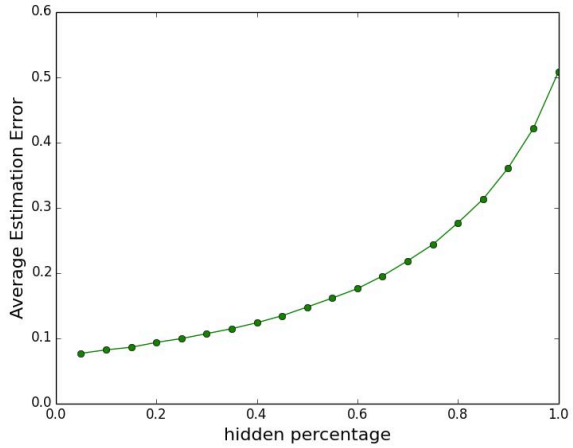


Fig. 5: Average estimation errors for the recombination model. x-axis is the percentage of hidden SNVs, and y-axis is the average estimation error in the test dataset.

	Variance	
	Component 1	Component 2
SyntheticData (M0)	0.923	0.837
SyntheticData (M1)	1.415	1.245
SyntheticData (M1-LD)	1.402	1.315
SyntheticData (M2)	1.841	1.551
SyntheticData (M3)	2.042	1.863
SyntheticData (M4)	2.171	2.113
SyntheticData (M5)	2.362	2.158
SyntheticData (M6)	2.402	2.351
SyntheticData (M7)	2.450	2.490
SyntheticData (RecombModel)	2.743	2.682
RealData (Training)	2.675	2.753
RealData (Test)	2.885	2.656

TABLE I: Variances of different models on the two principal components. The recombination model produces synthetic data that has a highly similar variance to that in real data, whereas other Markov models' performance improves as the order increases. Though not the best, the 6^{th}-order or 7^{th}-order models provide a reasonable simulation of real data.

components and projected different datasets on these two components to show how different models comply with the real data. Figure 6 provides the 2D visualisation of the synthetic sequences generated using the 1^{st}-order Markov model, the recombination model, and the 1^{st}-order Markov LD model as well as the training and the test data.

As Figure 6 shows, the synthetic genome sequences generated by the 1^{st}-order model built on the training dataset and the LD model built on the public AF and LD data have quite similar distributions as they are both based on 1^{st}-order Markov chain method. This demonstrates that publicly available genome sequences can be used as training data to model genome sequences, even though the number of the publicly available genome sequences is quite small. Furthermore, Figure 6 demonstrates that the synthetic data being generated by the recombination model has closer distribution to the training and test data, compared to the distribution of the synthetic data generated by 1^{st}-order model built on the training dataset and the LD model. This illustrates the effect of higher-order modelling as the genetic recombination model considers all the correlations between SNVs and builds a higher-order model.

For better understanding of the effects of higher-order modelling, we further used our training dataset to build the k^{th}-order Markov models for $k = 2, 3, 4, 5, 6, 7$ and generated 100 synthetic sequences using each model. Using PCA, we then mapped the synthetic sequences generated using the 6^{th}-order Markov model, the 7^{th}-order Markov model, and the recombination model, as well as the training and the test data which can be seen in Figure 7. As it shows, synthetic data generated by using higher-order Markov models have closer distribution to the distribution of the training and test data and hence with higher-order Markov models, the characteristics of the genome sequences can be better represented. But still, even the 7^{th}-order Markov model is far not enough to represent a real genomic data model, compared to the performance of a recombination model. A numerical comparison of the variances of different models on the two principal components are shown in Table I.

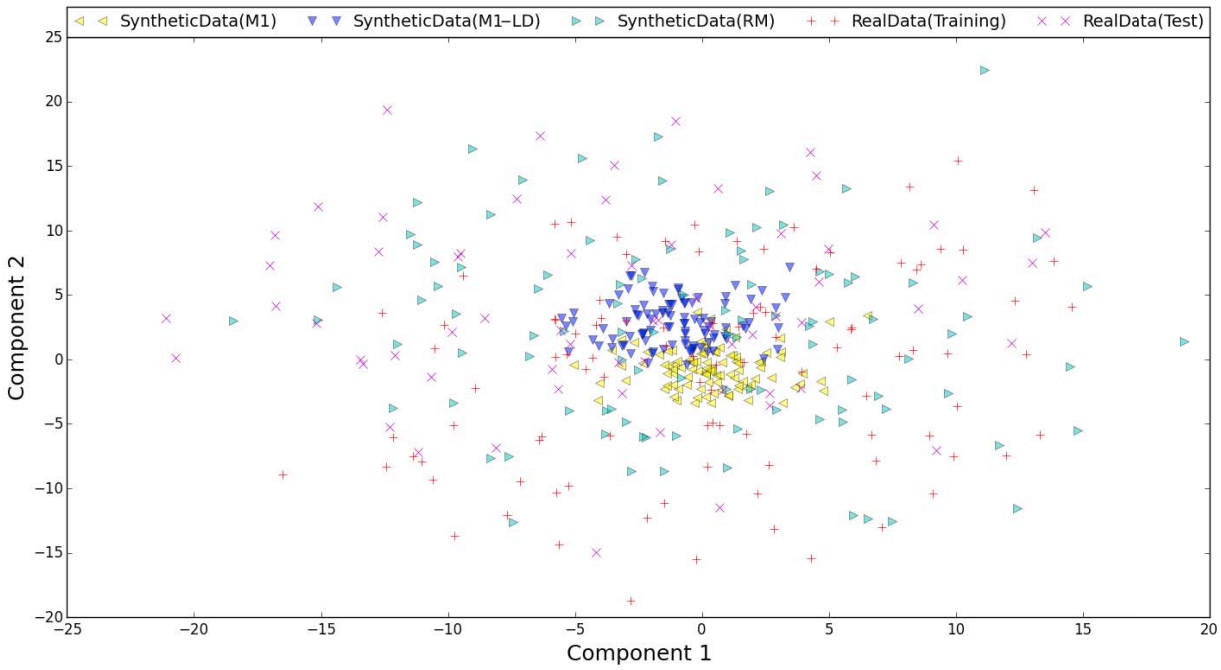


Fig. 6: 2D visualisation of the synthetic sequences generated using the 1^{st}-order Markov model (M1), the recombination model (RM), and the 1^{st}-order Markov LD model (M1-LD) as well as the training and the test data.

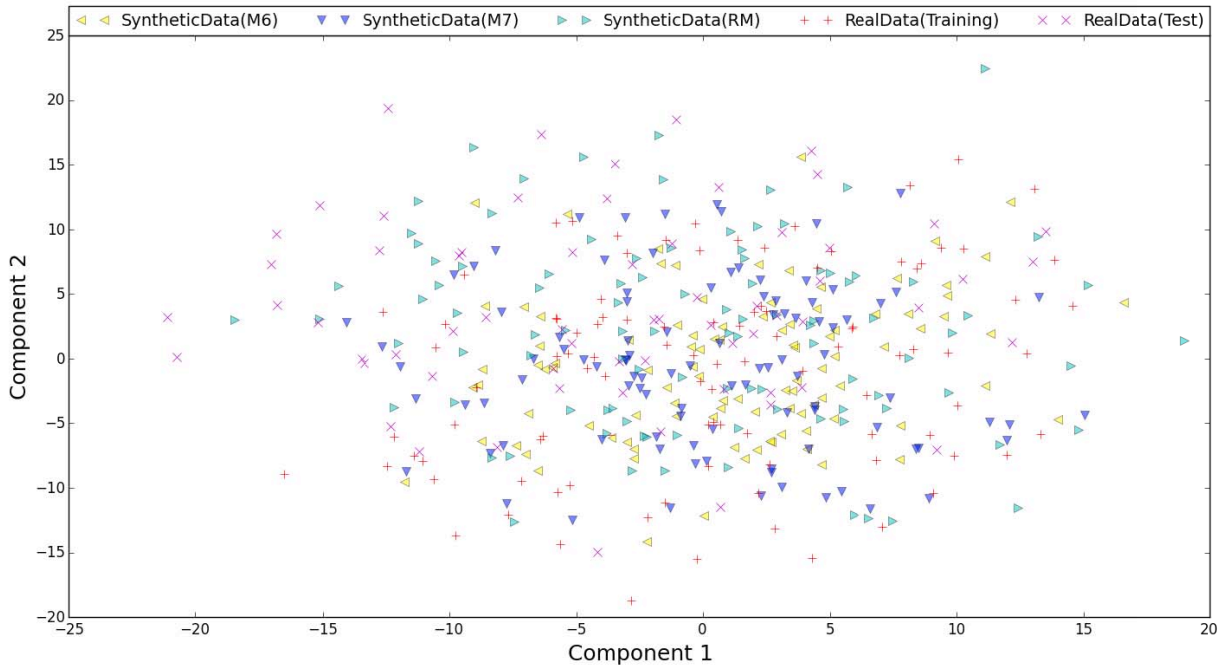


Fig. 7: 2D visualisation of the synthetic sequences generated using the 6^{th}-order Markov model (M6), 7^{th}-order Markov model (M7), and the recombination model (RM), as well as the training and the test data.

V. Discussion

Although higher-order models are preferred for the inference attack, directly building the models on a genotype dataset (Section III-A2) is not a scalable solution. As the order grows, the number of conditional probabilities that need to be estimated expands exponentially. Since the available genotype dataset is not large (165 samples), it might lead to an overfitting model if the order is too high, and thus the model would have a low statistical significance. In this work, we build the Markov models up to order 7, which is quite a high order considering the size of the genotype dataset. But with the recombination model, the number of parameters to be estimated is only linear with the number of SNVs. Indeed, only the recombination rates between every two adjacent SNVs need to be estimated. Hence, the recombination model is the genomic data model that should be used in practice, rather than the k^{th}-order model built on a genotype dataset. For a proof of concept, we discuss and use them together in this paper so that it is straightforward to observe how the model's order influences genomic privacy.

Our results build bridges between the diverse practices of evaluating genomic privacy in a genetic data sharing scenario. While some studies use pairwise LD to build models for such an evaluation ([4], [5], [6]), there is already practice in the literature where the recombination model is used for inferring genotypes that should not be released, like James Watson's ApoE gene status [12]. We have shown that pairwise LD information is not sufficient to quantify the sensitive information that will be leaked by such data releases once the SNV data is shared; indeed, there is a gap between the quantification with pairwise LD and the recombination model, which shows the large amount of sensitive information contained in the high-order SNV correlation. It gives researchers the potential to improve existing methods for the evaluation of genomic privacy under different scenarios. For instance, Humbert et al. [6] propose a graphical model to quantify kin genomic privacy in a genetic data sharing scenario that involves family members. By approximating high-order SNV correlations, the work shows an improvement of inference accuracy when the model integrates multiple pairwise LDs for each SNV. As there is no evidence that such an approximation with multiple pairwise LDs is sufficient, it seems plausible that we could improve on their results by combining their graphical model with the recombination model. Our future work will focus on the integration of high-order SNV correlation into the measurement of genomic privacy in various scenarios, including kin genomic privacy, in order to make more realistic and comprehensive assumptions about the adversary's power.

VI. Related Work

Much prior research has identified privacy breaches in genomic data, but most of this has relied on low-order SNV correlation. After Homer et al. [17] published their results on inferring individuals' contribution in genomic research, people become more concerned about the privacy leakage from a variety of analytical outputs from genomic research, including minor allele frequencies, χ^2-statistics and p-values; many such outputs were even removed from open-access databases. Even though Homer's attack relies on certain assumptions that might not hold in practice [18], such as the acquisition of the target's genome sequence and independence between SNVs, it does highlight the potential privacy threats arising from publishing genomic data and genomic computation results. A subsequent line of studies refined the above attack and thus unveiled yet more vulnerabilities of genomic data, [19], [20], [21]. Wang et al. [4] propose a more powerful re-identification attack by making use of p-values and linkage disequilibrium, which indicates that the privacy threat is more serious than what is shown by Homer's attack. Ayday et al. [5] quantify the privacy loss in a scenario of privacy-enhancing medical test and personalized medicine due to the inclusion of pairwise LD. Humbert et al. [6] propose a strong genotype inference attack by making use of both familial relationship and pairwise LD. Erlich and Narayanan [22] provide a comprehensive review about existing genetic privacy breaching techniques, including identity tracing attacks, SNP inference attacks and attribute disclosure attacks. Moreover, they show how an adversary can use the result of one attack as the input of a further attack, which chains the attacks as a complete pipeline, worsening the privacy breach that might arise from a single attack. Our work in this paper further explores the privacy problems that arise when people publish their genomic data online by comparing the inference performance based on different orders of correlation. The higher-order SNV correlation provides more information than the first-order correlation for an adversary and therefore represents a more realistic representation of an informed adversary.

The scenario we consider in this paper is probably best exemplified by the case of James Watson hiding his ApoE gene information that has been shown to be associated with Alzheimer's disease, when he shared his sequenced genome in public databases. Nyholt et al. [12] show that such gene information can be accurately estimated with the help of well-established genotype imputation techniques that use linkage disequilibrium and other released SNVs. Their study shows the difficulty of concealing SNVs in genomic data sharing. We provide a review on existing genomic data models of different Markov orders and discuss their implication for privacy under inference attacks.

VII. Conclusion and Future Work

Different genetic data models provide different levels of inference power for an adversary, depending on the order of SNV correlation that is captured by the models. Starting from a 0^{th}-order Markov chain (namely, assuming independence of SNVs), we show how the inference power gradually improves as the order increases. Capturing the highest order of correlation, the recombination model provides the best accuracy, and it is what an informed adversary will probably use. Hence, to give a reasonable evaluation of the privacy situation of sharing genetic data, one should consider the high-order correlation to avoid underestimating the power of a potential adversary. In our future work, we will provide frameworks to incorporate the high-order SNV correlation in the quantification of genomic privacy in various data sharing scenarios, such as for GWAS research and kin genomic privacy. On the genomic data protection aspect, we will look to develop privacy-preserving solutions that are robust against adversaries using high-order SNV correlations.

REFERENCES

[1] S. Jha, L. Kruger, and V. Shmatikov, "Towards practical privacy for genomic computation," in *IEEE Symposium on Security and Privacy*, 2008, pp. 216–230.

[2] M. Blanton, M. J. Atallah, K. B. Frikken, and Q. Malluhi, "Secure and efficient outsourcing of sequence comparisons," in *Computer Security–ESORICS*. Springer, 2012, pp. 505–522.

[3] C. A. Cassa, R. A. Miller, and K. D. Mandl, "A novel, privacy-preserving cryptographic approach for sharing sequencing data," *Journal of the American Medical Informatics Association*, vol. 20, no. 1, pp. 69–76, 2013.

[4] R. Wang, Y. F. Li, X. Wang, H. Tang, and X. Zhou, "Learning your identity and disease from research papers: Information leaks in genome wide association study," in *Proceedings of the 16th ACM conference on Computer and communications security*, 2009, pp. 534–544.

[5] E. Ayday, J. L. Raisaro, J.-P. Hubaux, and J. Rougemont, "Protecting and evaluating genomic privacy in medical tests and personalized medicine," in *Proceedings of the 12th ACM workshop on Workshop on privacy in the electronic society*, 2013, pp. 95–106.

[6] M. Humbert, E. Ayday, J.-P. Hubaux, and A. Telenti, "Addressing the concerns of the Lacks family: Quantification of kin genomic privacy," in *Proceedings of the ACM SIGSAC conference on Computer & communications security*, 2013, pp. 1141–1152.

[7] R. Gorelick and M. D. Laubichler, "Decomposing multilocus linkage disequilibrium," *Genetics*, vol. 166, no. 3, pp. 1581–1583, 2004.

[8] Y. Kim, S. Feng, and Z.-B. Zeng, "Measuring and partitioning the high-order linkage disequilibrium by multiple order markov chains," *Genetic epidemiology*, vol. 32, no. 4, pp. 301–312, 2008.

[9] S. Feng and S. Wang, "Summarizing and quantifying multilocus linkage disequilibrium patterns with multi-order markov chain models," *Journal of biopharmaceutical statistics*, vol. 20, no. 2, pp. 441–453, 2010.

[10] M. Humbert, E. Ayday, J.-P. Hubaux, and A. Telenti, "Reconciling utility with privacy in genomics," in *Proceedings of the 13th Workshop on Privacy in the Electronic Society*. ACM, 2014, pp. 11–20.

[11] N. Li and M. Stephens, "Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data," *Genetics*, vol. 165, pp. 2213–2233, 2003.

[12] D. R. Nyholt, C.-E. Yu, and P. M. Visscher, "On Jim Watson's APOE status: genetic information is hard to hide," *European Journal of Human Genetics*, vol. 17, no. 2, p. 147, 2009.

[13] https://opensnp.org/, [Online; accessed 6-January-2015].

[14] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, pp. 257–286, 1989.

[15] J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly, "A new multipoint method for genome-wide association studies by imputation of genotypes," *Nature genetics*, vol. 39, pp. 906–913, 2007.

[16] http://hapmap.ncbi.nlm.nih.gov/downloads/index.html.en, [Online; accessed 6-January-2015].

[17] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig, "Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays," *PLoS genetics*, August 29, 2008.

[18] R. Braun, W. Rowe, C. Schaefer, J. Zhang, and K. Buetow, "Needles in the haystack: identifying individuals present in pooled genomic data," *PLoS genetics*, vol. 5, no. 10, 2009.

[19] K. B. Jacobs, M. Yeager, S. Wacholder, D. Craig, P. Kraft, D. J. Hunter, J. Paschal, T. A. Manolio, M. Tucker, R. N. Hoover *et al.*, "A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies," *Nature genetics*, vol. 41, no. 11, pp. 1253–1257, 2009.

[20] P. M. Visscher and W. G. Hill, "The limits of individual identification from sample allele frequencies: theory and statistical analysis," *PLoS genetics*, vol. 5, no. 10, p. e1000628, 2009.

[21] S. Sankararaman, G. Obozinski, M. I. Jordan, and E. Halperin, "Genomic privacy and limits of individual detection in a pool," *Nature genetics*, vol. 41, no. 9, pp. 965–967, 2009.

[22] Y. Erlich and A. Narayanan, "Routes for breaching and protecting genetic privacy," *Nature Reviews Genetics*, vol. 15, no. 6, pp. 409–421, 2014.