SUPPLEMENTAL METHODS for “Proxy Panels enable Privacy-aware Outsourcing of Genotype Imputation”

Degui Zhi[1], Xiaoqian Jiang[2], Arif Harmanci[1,2,*]

[1] Department of Bioinformatics and Systems Medicine, D. Bradley McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA.
[2] Department of Health Data Science and Artificial Intelligence, D. Bradley McWilliams School of Biomedical Informatics, University of Texas Health Science Center, Houston, TX, 77030, USA.
[*] Corresponding Author: Arif Harmanci, Email: Arif.O.Harmanci@uth.tmc.edu, Phone: 713-500-3650

Table of Contents

On Privacy Concerns in Genetic Research and in Genotype Imputation-related Tasks

Arguably, protection of privacy of participants and families is one of the most challenging aspects of genetic data sharing. These ethical challenges are multidimensional and complex [1], e.g., usage for forensic purposes[2–4], and concerns about discrimination[5,6]. Public is becoming increasingly aware of the ethical and privacy implications around sharing genetic data, and regard privacy as one of the top factors while deciding who they share genetic data[7,8]. Numerous negative experiences, e.g., Havasupai Tribe[9,10], may undermine the trust in the research community due to lack of ethical reporting standards and insensitivity to cultural norms. Due to these concerns, datasets are often siloed in the central servers and are protected by data usage agreements and severely hamper collaborative research opportunities[11,12] even within the same consortia[13,14]. Although legislation such as GDPR in the EU and HIPAA[15] in the USA aim at protecting personal data, their interpretation is not clear about genetic data. For example, GDPR requires using new technological methods to make sure that the data is "irreversibly de-identified" unless consent is available for sharing identifiable data. However, even when consent is granted for sharing genomes, individuals may not consent to the downstream characterization of their genomes[16].

The main concerns around genomic privacy[17–20] are related to the re-identification of an individual by "singling-out" (GDPR) within a cohort. In the context of imputation outsourcing, the imputation server (which are sometimes referred to as the "processors"), and the data owners/controllers (query and reference data owners) must make sure to take the precautions to minimize the risks and evaluate them effectively. This is a very challenging process while individual level datasets (from both the client and the reference panel owners) are being shared with the imputation servers.

Due to the high dimensional and complex correlative structure of genetic data [16,21,22], many routes are open for re-identification of individuals[19]. For example, only 100 carefully selected SNPs are sufficient to identify any individual[23]. The most famous attacks, including Homer's t-statistics[24], Sankararaman's LRT[25] demonstrated that an individual's participation in a genome-wide association studies (GWAS) study can be reliably identified even when only allele frequencies of the study are known. These attacks were alarming at the time because of their simplicity and their applicability to summary statistics[26,27]. Further studies identified that participation in different studies can lead to linking attacks and reveal sensitive information[28,29] of participants and relatives[30,31]. Other attacks that make use of haplotype frequencies have been proposed to identify participation[32,33]. Recent studies demonstrated that genetic beacons[34] (where the existence of variants in a database are queried as yes/no answers) are vulnerable to Bustamante's attacks[35] and more advanced attacks[36,37] where an adversary can identify participation of an individual in a beacon by adversarial querying, which may impact biobank scale databases. For example, the knowledge of the rare variant positions can be used to reveal participation in TOPMed panel using results obtained from TOPMed imputation server. One of the important factors that make these attacks possible is that they rely on a naive data-sharing model where raw data is made available to entities without a specific focus on a task such as genotype imputation. Furthermore, several studies showcased that the most well-known attacks are applied in very well-controlled scenarios and they lack of formal treatment of the false positive rates [38] when the assumptions are not satisfied [39].

By far, the most popular genetic data sharing model is the restricted access model that model relies on users signing agreements for each new dataset. Differential privacy[40,41] (DP) approaches design privacy-preserving data release mechanisms where aggregate statistics from data are computed and noise is added to the released statistics to enable privacy. While DP provides strong privacy guarantees, the noise addition is a major drawback since a naive noise addition may corrupt certain data types (such as genotype data) to the extent that the data become unusable. Encryption-based approaches are the most rigorous route to securely share genetic data albeit with low practicality and are applied to the field of genotype imputation[42,43], and association studies [44–47], database sequence queries [48], and read mapping[49,50]. Methods such as homomorphic encryption[51,52] enable analyzing and processing encrypted data directly. There are, however, a substantial computational (requirement for reformulation of algorithms under cryptographic primitives [53]), storage (ciphertext expansion), and maintenance costs. Moreover, each algorithm must be re-implemented and optimized to efficiently analyze the encrypted data. Multiparty computation (MPC) based systems [54,55] provide cryptographic security by dividing data among multiple parties. MPC-based systems rely on heavy communication and computational requirements on the parties [46] and require algorithms to be redesigned with careful reformulation and parameter selection (circuit depth, ciphertext size) to ensure that algorithms can be executed efficiently. Consequently, HE and MPC methods are not widely adopted in the community, yet. Trusted execution environments (Intel SGX, AMD SEV) use security-by-engineering which are vulnerable to numerous attacks[56,57] and rely on trust on a corporate entity. Of note, Intel SGX is currently deprecated for client-side processors. Although several anonymization techniques exist[58,59], they require reformulation of specific tasks (e.g., relative searching) and may not be easily applicable to different high throughput tasks.

Synthetic datasets have been explored as a potential approach that can be useful for protecting privacy[60] in collaborative studies. In these approaches, synthetic datasets are generated at each collaborating entity. In this framework, the entities generate a representative synthetic dataset using their local sensitive datasets. In the analysis, synthetic datasets are used instead of the original sensitive datasets. The privacy of participants is protected since the sensitive datasets are never used in analysis. In genomics, synthetic datasets are used for protecting reference panels in genotype imputation[61,62], and for analysis of ancestral simulations[63,64]. While these approaches provide means for generating synthetic datasets, privacy is still not explicitly introduced into the synthetic data generation models. For example, RESHAPE[61] uses a sampling procedure to build a “mosaicized” panel starting from the reference datasets. RESHAPE assumes to provide privacy to the subjects in the reference panel, which can be shared publicly and can be used as imputation panels. However, the allele frequencies and coordinates for the variants at the rare and ultra-rare category (including singletons that leak immediate membership information) are preserved in the synthetic data, which may make the synthetic datasets vulnerable to well-known re-identification attacks. Given the dominating number of rare variants in datasets, censoring or budgeting these variants does not effectively protect the data[35,65], and would severely decrease utility since most of the variants would be censored. There are currently methodological gaps that can provide flexible privacy protections while also enabling efficient calculations.

Impact of Proxy Generation Parameters on Imputation Accuracy

We studied the impact of proxy panel generation parameters on the imputation protocol accuracy. For this test, we randomly divided the 661 subjects into two panels (330 and 331 subjects) in African population (AFR) of The 1000 Genomes Project, and extracted the 16,114 variants on chromosome 22 of Illumina Duo v3 array to be used as the typed variants. We imputed the 201,040 untyped variant genotypes on first AFR panel using the second AFR panel as the reference panel. Each parameter was varied while other parameters are held constant at default values (Methods). Imputation accuracy is estimated using genotype level R2 metric.

Hash Weight Selection. We first varied the parameters for random weight selection for 1st, 2nd, and 3rd degree components of the rolling hash. Among these, 1st degree components had the most impact on the accuracy while 2nd and 3rd degree component selection probabilities did not change the accuracy substantially (Supplemental Fig. S1a, 1b, 1c). We chose to keep 2nd and 3rd degree weights to increase complexity of the hashing function. We also observed that the minimum number hash weights did not make a substantial difference in accuracy (Supplemental Fig. S1d). Interestingly, we observed that increasing the variant selection weight parameter ($N_e^{(w)}$), which increases probability of using more local variants for hashing each variant, decreases imputation accuracy (Supplemental Fig. S4e). This result suggests that spreading the haplotype information across hashed alleles (lower variant selection parameter) increases the accuracy of imputation. The rolling hash window size (n_{vic}) also has a strong impact on accuracy (Supplemental Fig. S4f). We observed that increasing hashing window size increased accuracy up to 13 variants ($n_{vic} = 6$) and accuracy decreased at 15-variant vicinity. This suggests that spreading the allelic information on longer windows can increase accuracy. As expected, when the hashing window length becomes too large, the LD information in proxy reference panel may become spread adversely for accurate imputation.

Variant and Genetic Map anonymization Parameters. We assessed the two parameters that are used for anonymizing the genetic map distances (σ_g) and untyped variant permutation filter window length. We observed that increasing σ_g decreases accuracy (Supplemental Fig. S1g), especially between $\sigma_g = 0.01$ and $\sigma_g = 0.05$. The untyped variant permutation window size did not make a substantial difference in accuracy (Supplemental Fig. S1h). This supports our expectation that HMM-based imputation tools are insensitive to the permutation of untyped variants when they are permuted between the surrounding typed variants.

Comparison of Proxy Panels with Original Panels

In this section, we compare the proxy panel and the original panel with respect to their simple statistics. Since entities would have access to only the proxy panel, the comparisons presented here would not be possible by an adversarial honest-but-curious entity, which is the main adversarial model we are focusing on. We, however, compare the original panel and proxy panels to highlight how well the proxy panels reflect the original panels that they were derived from.

Each haplotype in the proxy haplotype panel is generated by 1) resampling, 2) hashing of the typed variants alleles and 3) partitioning and permutation of the untyped variants (only for the reference site), 4) Protection of the variant positions and genetic maps. We first asked whether the proxy haplotypes leak information about the original panel at the level of variants. We compared the proxy panel and the original panel in terms of genotype and variant-variant linkages.

Comparison of genotypes and haplotype frequencies in original and proxy panels. We first assessed the correlation between the original variant genotypes. We extracted the genotypes of 661 individuals in AFR population of the 1000 Genomes Project and divided the panel into 2 subpanels comprising 330 and 331 subjects. We used the first panel (original panel) and generated the corresponding proxy panel with the same number of subjects. The second panel is used as a matching control (holdout) panel. For each variant, we correlated the original panel genotypes (alternate allele dosage) with the genotypes of the proxy panel among individuals. We observed no significant correlation between the genotypes of the original and the proxy panel (Supplemental Fig. S4a). Further, the genotype correlations of original-vs-proxy did not exhibit any enrichment compared to the correlations of original-vs-holdout panel. This is expected because the resampling step breaks one-to-one correspondence between subjects. When we remove resampling step in proxy panel generation (Fig. Supplementary S4a), the distribution of correlations exhibits a large variance. This is also expected since for some variants, the hashing may involve a small number of surrounding variants due to random hashing weight selection (Methods) which leads to high correlation to the original variant alleles.

We next compared the allele frequency of each variant in the original and proxy panels. We observed there is virtually no correlation between allele frequencies of variants in the original and proxy panels (Supplemental Fig. S4b). As expected, holdout panel exhibits very high concordance to the original panel in allele frequencies. (Supplemental Fig. S4c)

Variant-variant Linkage and Local Haplotype Frequency Concordance between Original and Proxy Panels. We next calculated the Pearson R2 coefficient between genotypes of variant pairs in a sliding window of 500 variants to estimate the linkage patterns in original and proxy panels. We observed that the variant-variant correlation patterns do not exhibit clear concordance although there is significant correlation of (Pearson R=0.33) (Supplemental Fig. S4d). It should be noted that the linkage concordance between the original and control panels is much clearer (Supplemental Fig. S4e) (Pearson R=0.99). We next compared the local haplotype counts (haplotype comprising k-typed variants) between the original, proxy, and the control panels by extracting the k-mers (k=11) in sliding windows and calculating the number of unique k-mers in each window. Although there is a small correlation between the number of unique k-mers at each position of the proxy and original panels (Pearson R=0.01), the concordance is much weaker (Supplemental Fig. S4f, S4g) than the two matching panels (Supplemental Fig. S4h, S4i). In particular, we observed that the proxy panel harbor an excess of unique k-mers per position compared to the original panel. This result demonstrates that random hashing increases the entropy of the k-mer (local haplotype) distribution at each position.

Population Stratification of Proxy-Panels. We finally performed Principal Component Analysis (PCA) using the proxy panel genotypes to evaluate how well the population-level information is reflected in proxy panels in comparison to the original panel. In this analysis, we first extracted the genotypes of the 1000 Genomes Project from the EUR, AFR, and EAS populations and performed PCA on the original genotypes and the proxy genotypes. When the sampling step is performed, PCA did not provide any information about the ancestry of the proxy panels (Supplemental Fig. S4j). When we generated the proxy panel without the resampling step, the proxy panel separated into 3 distinct clusters. This is expected because sampling shuffles the haplotypes into mosaics and ancestry information is lost among the sampled mosaic haplotypes. To assess if it is possible to match the subjects to each other when sampling is not used, we aligned the SNPs by their chromosomal order and we merged proxy and original genotype matrices into a pooled genotype matrix and performed PCA on the pooled matrix. This shows that the original and proxy panels separate into 6 distinct clusters (Each corresponding to an ancestral group in original and proxy panels) with no clear mixing between them.

Overall, these results indicate that proxy panels do not leak substantial first order (allele frequencies) or second order information (variant linkage or population stratification) about the original sensitive panels and it is unlikely to perform re-identification using proxy panels without a malicious intent (e.g., stealing hashing parameters). Application of Homer's t-statistic and LRT attacks are not immediately possible on proxy panels since variant coordinates are obfuscated and the genotypes are hashed. We also hypothesize that linking attacks (matching each subject in proxy panel to an external panel) are unlikely when sampling step is used because each proxy-haplotype is a random mosaic of the original haplotypes.

On Protection of Variant positions and Genetic maps by ProxyTyper

Homer(Homer et al. 2008) and LRT-type re-identification attacks(Sankararaman et al. 2009) attacks require matching the variant positions between the reference panel and the mixture panel so that attack statistics can be calculated for identifying a target individual with known genotypes. To ensure that the proxy variant positions are not directly linkable to external panels, ProxyTyper obfuscates the variant positions by replacing variant positions to be distributed uniformly on an anonymous chromosome with length of, by default, 10^8 nucleotides. Even though the variant positions are obfuscated, genetic maps can be aligned to public sources to match them and reveal exact locations. To protect the genetic maps that are shared among the sites, ProxyTyper perturbs the maps by noise addition with a user-specified deviation σ_{map}, measured in centiMorgans (cM). For example, given $\sigma_{map} = 0.05\ cM$, corresponds to approximately 50 kilobases (1 megabase/cM) of perturbation to the position of the variant. ProxyTyper generates anonymized genetic maps only for the typed variants (other entries in the original map are discarded) because untyped variant genetic maps are interpolated by imputation software from typed variants. It should be noted that imputation tools do not require variant coordinates when genetic maps are available. This is expected since imputation HMMs work on executing Li-Stephens model using the genetic distances and coordinates (in base pairs) are not needed. The coordinate anonymization for typed and untyped variants is basically a one-to-one mapping between original and anonymized coordinates. ProxyTyper stores these in extended BED files. The anonymized genetic maps for typed variants (in anonymized coordinates) are stored in Plink formatted genetic distance map files.

The main objective of coordinate and genetic map anonymization mechanism is protection at a local scale, i.e., local perturbation of the variants to make sure variants cannot be exactly pinpointed. Note that this mechanism does not protect the coordinates at a global manner since even the total genetic length (in centiMorgans) of the chromosomes leak the chromosome identity. After assigning the chromosome, approximate positions can be assigned by aligning the noise genetic distance to publicly available genetic maps. Although the public genetic maps can be used for roughly aligning and de-anonymizing the proxy panel coordinates, it is not likely to perform a perfect match.

Steps of ProxyType'd Genotype Imputation Protocols

Phased Query Protocol (Fig 3b). Given a phased panel, the query initiates the protocol by sending the typed variants to the reference site (e.g., a BED file). The reference site initializes the parameters for the augmentation, hashing, permutation, and coordinate anonymization mechanisms and sends them to the query site. Sites then execute following mechanisms in the order:

1) Resample haplotypes: Resampling is performed independently on query and reference sites. This step does not require communication between the sites or the mechanism parameters. The resampled panel size is selected based on the underlying panel size.

2) Augment, hash, permute typed variants: Typed variants are first augmented. By default, ProxyTyper performs 3 recursive augmentations and augments each typed variant within 2 tags in the vicinity. Variants are hashed using 11 typed variant vicinity and permuted within 7 typed variant vicinity where in each window is permuted with 20% probability.

3) Anonymize coordinates and genetic maps for all variants: The coordinate system is anonymized to a chromosome with length 100 megabases. The genetic map is anonymized with the default noise level.

4) Untyped Variant Partitioning at the Reference Site: The reference site generates the parameters for the untyped variant partitioning mechanism and stores them locally.

5) Imputation on the Server: The query and reference proxy panels and the anonymized genetic map are sent to the imputation server (Fig. 3b), which executes Beagle and sends the imputed panel for the proxy query panel back to the query site. Note that although the query panel is phased, it is sent as an unphased panel to the imputation server.

6) Local Re-Imputation on Query Site: As the query proxy panel was generated starting from the resampled query panel (Step 1 above), the imputed genotypes do not correspond to the original query subjects. To map the imputed genotypes back to the original panel, the query panel performs Steps 2 and 3 on the original query panel (without resampling step) and use this encoded panel as input to perform a local re-imputation. In the local re-imputation, the panel downloaded from the imputation server is used as the reference panel option.

The phased protocol can be simplified using two subprotocols in Fig. 3a, which demonstrates the mechanisms can be used as building blocks of different subprotocols and larger protocols. Note that These steps are implemented into a complete phased query protocol in ProxyTyper's codebase with the default parameters described above.

Unphased Query Protocol (Fig 3c). When the query panel is unphased, haplotype resampling is used only on the reference site. To protect typed variants, sites use augmentation (by default 4 recursive augmentations) followed by permutation mechanisms (by default 9-typed variant long permutation windows with 50% permutation on each window). Coordinates and genetic maps are anonymized as phased protocol. The reference site on the resampled panels uses the untyped variant partitioning to protect untyped variation positions and genotypes. Since the query site does not use resampling in the unphased query protocol, The imputed panel returned from Imputation Server is used as it is, which renders the unphased query protocol simpler than the phased query protocol, offering a simpler protocol.

From participant privacy perspective, phased query protocol is more privacy preserving than the unphased protocol since it uses resampling on the query panel. Further, the query and reference site alleles are hashed, which adds another layer of protection. Unphased query protocol relies mainly on augmentation followed by a stronger permutation mechanism.

Practical Attack Routes to Proxy Panels generated by ProxyTyper

No data protection scheme is perfectly secure, even when the encryption/hashing keys are protected. We therefore strongly recommend protecting the proxy panels from public access.

We discuss a pipeline that we would use to attack a proxy panel generated by ProxyTyper. We assume that the adversary has following:

1) Access to public resources such as 1000 Genomes Project, dbSNP, UCSC Genome Browser, etc.
2) Gained access to a proxy panel built by ProxyTyper's protocol,
3) Adversary is incentivized to decode proxy panels.

Selection of Attack Reference Panel: The most feasible attacks would start with an "attack-reference panel" because the genomic coordinates are not available to the adversary. By far, the most feasible panel for this is the 1000 Genomes Project panel, which is public.

Coordinate Matching by Aligning Anonymized Genetic Maps to Public Genetic Maps: Next, the coordinates must be matched between the attack-reference panel and the proxy panel. It would be unlikely to accurately do a 1-to-1 mapping of the variants in the proxy panel to the variants in the attack-reference panel because ProxyTyper uses typed variant augmentations and randomized local permutations. Thus, the most practical way to perform matching is where the attacker will map approximately where each proxy variant is on the attack-reference panel by aligning the anonymized genetic maps to a publicly available genetic maps (e.g., HAPMAP's genetic maps).

Frequency Analysis on the Haplotype Frequencies: Finally, the reference panel is compared with the proxy panel with a *frequency analysis* (the standard attack mode for ciphertext-only-attacks, i.e., when only ciphertext is available to an attacker or cryptanalyst) to decode the hashed alleles using decoding HMM, which has at least the same computational complexity as imputation. Note that, in this setup, what the adversary decodes are not the variants in the proxy panel but the variants in the attack-reference panel. This is because the coordinate mapping from attack-reference to the proxy panel does not reveal exact mappings of each variant to the adversary. In other words, the adversary could only decode which variants they have in the attack-reference panel. This has certain implications on the risks and design of future protection mechanisms:

1) The decoded variants would not reveal the actual variants in the sensitive panel; they would only reveal the variants on the attack-reference panel. One could argue that this is a different type of smaller risk of leakage since actual sensitive variants are still not directly accessible by the adversary, i.e., the adversary decodes a "projection" of the sensitive panel on the attack-reference panel.
2) Any deviations between the attack-reference panel and the original sensitive panel will lead to differences in haplotype frequencies and will lead to low decoding accuracy. This means that resampling mechanisms can aim at slightly shifting haplotype frequencies such that public panels are not similar in haplotype frequencies to the resampled panels. The simplest strategy to accomplish this is to "dope" a small fraction of randomly selected haplotypes from different ancestries (e.g. from the 1000 Genomes Project haplotypes) while resampling.
3) The strategy in (2) also makes it more challenging to perform the membership inference attacks (e.g., Homer et al. t-test) since the resampled panels are diverged from the publicly available reference panels.

Below, we present a decoding HMM that performs a more complete frequency analysis by taking into account the continuity of the haplotypes and test its accuracy on a well-controlled dataset.

Viterbi Decoding of Attack on Typed Variants and Likelihood Ratio Test Reidentification

We describe the Viterbi decoding algorithm that is used for decoding the hashed alleles in the proxy panels. Decoding algorithm is similar to an imputation HMM that decodes the hashed (proxy) alleles in a proxy panel by combining a frequency analysis of proxy panel k-mers with reference panel k-mers. It should be noted that the k-mer length in decoding is independent of the proxy panel generation k-mer lengths.

Inputs: The adversary has access to a proxy panel that is being decoded. Adversary uses a public reference panel to estimate the frequencies of the cleartext k-mers and match them at each k-mer. It should also be noted that the proxy panel's variant coordinates are anonymized, i.e., adversary needs to predict the positions of the typed variants. Rather than predicting the positions of variants, we assumed that the attacker approximately predicts where the variants are using, for example, LD statistics. To simulate this, we divided the 87,960 typed variants into 2 equal sized sets by assigning every 1st variant to the first set and every 2nd variant to the second set, i.e., the variants sets are interleaved with respect to each other. Coordinates of the 1st variant set are used for the proxy panel and those of the 2nd set is used for the reference panel. Simply put, this corresponds to the attacker using an interleaved set of variants for the attack. The proxy panel is generated from a sample of 100 subjects from the AFR populations in the 1000 Genomes Project. For decreasing computational complexity, we focused on a region of 100 megabases (chr1:10,000,000-110,000,000), which contains 19,379 variants. The reference panel (in comprises the genotypes of 461 subjects. We also used a hold-out panel of 100 subjects that are used as a control panel.

Transition and Emission Probabilities: The decoding HMM uses a modification of the Li-Stephens model[66] that is used in imputation methods. The decoding HMM runs on the coordinates of the reference panel and uses each reference haplotype as a state in the Markov model. The transition probability at each position is assigned using the genetic distances and the consecutive k-mer concordance. At reference variant position i, the transition probability is defined as:

$$\tau_i^{(LS)}(j' \rightarrow j) = \begin{cases} p_{recomb} = \dfrac{\left(1 - \exp\left(-4 \times N_e^{(dec)} \times \dfrac{\left(\Delta_g[i] - \Delta_g[i-1]\right)}{N_{ref}}\right)\right)}{N_{ref}}, & j \neq j' \\ 1 - \left(N_{ref} - 1\right) \times p_{recomb}, & j = j' \end{cases}$$

which is described by Li-Stephens model. Note that the transition does not rely on the haplotype indices (i.e., states) except for the staying on the same state, i.e., all same state transitions are assigned the same probability. $N_e^{(dec)}$ is set to default value of 10^6. We modify this transition probability to introduce the concordance of haplotype frequencies:

$$\tau_i^{(HapFreq)}(j' \rightarrow j) = \tau_i^{(LS)}(j' \rightarrow j) \times \frac{\exp\left(-1 \times \left| f\left(\widetilde{H}_{[i-k,i+k],j'}\right) - f\left(H_{[i-k,i+k],j}^{(ref)}\right)\right|\right)}{\Xi},$$

where k denotes the vicinity length (i.e., total k-mer length is $(2k+1)$) and Ξ denotes the normalization factor over all haplotypes. Above equation integrates concordance of the allele frequencies between the proxy panel ($f\left(\widetilde{H}_{[i-k,i+k],j'}\right)$) and the reference panel ($f\left(H_{[i-k,i+k],j}^{(ref)}\right)$) to the transition score. We further need to introduce the constraint on consistency of the consecutive k-mers on the current haplotype (i.e., state j) state and previous state (state k) at variant position $i-1$:

$$\tau_i^{(Dec)}(j' \rightarrow j) = \tau_i^{(HapFreq)}(j' \rightarrow j) \times \exp\left(-\kappa_{conc} \times \frac{\left| H_{[i-k,i-1+k],j'}^{(ref)} - H_{[i-k,i-1+k],j}^{(ref)} \right|}{(2k+1)}\right)$$

where we score the distance between the overlapping substrings of the two k-mers at positions $[i-1-k, i-1+k]$ and $[i-k, i+k]$. The modifications in the equations constrain the concordance between the frequencies of the reference panel and the proxy panel, and also the concordance of the emitted reference panel k-mers between current haplotype and the previous haplotype. This way HMM can follow haplotypes with concordant consecutive k-mers by making transitions between them. κ_{conc} denotes a weight that tunes the weight of k-mer concordance between two positions. This value is set as:

$$\kappa_{conc} = -8.8 \times (2k+1)$$

which penalizes every additional mismatch between the consecutive k-mers by 10^{-4}, which is the default mismatch score in BEAGLE for allelic errors.

$\tau_i^{(Dec)}(i \rightarrow j)$ denotes the overall probability of making the transition from state (i.e., haplotype) i to state j given the query k-mer and the reference k-mers at the current typed variant position. Note that this transition probability takes into account (1) the recombination rates ($\tau_i^{(LS)}(j' \rightarrow j)$), (2) the k-mer frequency concordance ($\tau_i^{(HapFreq)}$), (3) Continuity of the k-mers by comparing the overlapping substrings in the k-mers ($\tau_i^{(Dec)}(j' \rightarrow j)$).

Scoring Matrix Calculation. Given a reference panel haplotype that consists of n_{var} variants, we allocate $n_{var} \times N_{ref}$ matrix of probability scores, denoted by S. We initialize the scoring matrix at the first variant position uniformly. At state (i.e., haplotype) j, the initialization is performed as:

$$S_{0,j} = \frac{1}{N_{ref}}$$

At variant position $i > 0$, the score at state (i.e., haplotype) j is calculated by evaluating all state transitions at the previous variant position $i - 1$:

$$S_{i,j} = \max \begin{pmatrix} S_{i,1} \times \tau_i^{(Dec)}(1 \rightarrow j), \\ S_{i,2} \times \tau_i^{(Dec)}(2 \rightarrow j), \\ \ldots \\ S_{i,N_{ref}} \times \tau_i^{(Dec)}(N_{ref} \rightarrow j) \end{pmatrix}$$

This recursion is calculated starting from $i = 1$ to $i = n_{typed}$ for all reference haplotypes $j < N_{ref}$. For each proxy haplotype, the scoring matrix is calculated and the final score matrix is traced back using Viterbi algorithm to identify the highest scoring haplotype sequence over the reference haplotypes. The Viterbi path is basically a path through the haplotypes of the reference panel. The final decoded allele sequence is obtained by using the allele in the middle of the k-mer window.

Reidentification of individuals within decoded proxy panel using LRT Attack. After the alleles are decoded, we use the decoded haplotype matrix as a pool in which a target individual with known genome is searched. The adversary uses the same reference panel used for performing the re-identification attack. The LRT attack statistic is calculated as described in Sankararaman et. al. (Supplementary Information page 25 in [25]) by comparing the allele frequencies of variants between the pool and the reference dataset. LRT statistic is calculated as:

$$L = \sum_{1 \le j \le M} \left[h_j \times \log\left(\frac{\tilde{p}_j}{p_j}\right) + \left(1 - h_j\right) \times \log\left(\frac{1 - \tilde{p}_j}{1 - p_j}\right) \right]$$

where $h_j \in \{0,1\}$ denotes the binary allele value at j^{th} variant, $\tilde{p}_j$ denotes the frequency of alternate allele ($h_j = 1$) in the pool and p_j denotes the frequency of alternate allele in the reference panel. LRT statistics are calculated for all decoded haplotypes.

We evaluated three case scenarios to calculate the re-identification statistics that are shown in Supplemental Figure S5:

1) Use decoded proxy panel as the pool (Actual attack scenario, Supplemental Fig. S5b): This is the basic scenario when proxy panel is used for protecting the panel. Adversary tries to identify if an individual with known genotypes is in the decoded panel.

2) Use the cleartext original panel as the pool (Baseline scenario with original data for pool panel with no protection, Supplemental Fig. S5c): This is the baseline attack scenario where adversary has access to the cleartext sensitive panel and tries to identify individual's participation in the panel.
3) Use the cleartext resampled original panel as the pool (Control case scenario with sampled mosaic panel without allele hashing as the pool panel, Supplemental Fig. S5d): This scenario evaluates the participation prediction when only resampling is used.

Application of the Viterbi Decoding of Proxy Alleles and Re-identification Attacks

Our previous results indicate that there is small but significant correlation linkage of consecutive variants in original and proxy panels (Supplemental Fig. S5a). The main risk for re-identification in the proxy-panels will stem from decoding of the proxy-alleles to infer the original panel by matching the haplotype frequencies using an external panel, followed by a re-identification attack such as an LRT attack. This requires an adversary to decode the hash function weights at every variant, which is unlikely without external knowledge (e.g., stolen or leaked parameters which may lead to *known-plaintext attacks*). We assume that the adversary only has access to the proxy panel haplotype matrix with some knowledge of the hashing function such as the size of vicinity window but does not know the weights of hashing. While it is not immediately available, the size of vicinity window used for hashing could potentially be inferred from the proxy panel's k-mer frequency statistics.

One attack for decoding the proxy panels can make use of two properties of the hash function: Firstly, hashing of consecutive variants use overlapping windows, i.e., if the adversary can guess the original k-mers at a position, these k-mers can be used to make a better decoding prediction for the next variant. Secondly, the k-mer frequencies between the proxy panel and original panel are somewhat concordant, e.g., a high frequency k-mer in the original panel is likely a high frequency k-mer in proxy panel and vice versa. Thus, the adversary does not have to infer the weights of the hash function but try to match the k-mers based on their similarity in proxy and reference frequencies to decode and map the proxy k-mers back to original k-mers while using the k-mer at previous position as a constraint and perform a *frequency analysis attack*[67]. This process can be implemented in an HMM (similar to genotype imputation) while considering the recombination rates (estimated from genetic distances) as further constraints on the decoded alleles.

To test the viability of this attack, we implemented a modified version of the Li-Stephens hidden Markov model that can decode the k-mers in proxy haplotypes by comparing their frequencies in the proxy panel to those in a separate reference dataset (Methods). While decoding a proxy haplotype, HMM starts from the first variant. At each variant on the proxy haplotype, HMM constrains the possible reference k-mers in comparison with the previous position, and the frequency concordance between the proxy haplotype's k-mers and the original k-mers, and integrates the recombination frequencies (i.e. haplotype switch) using the genetic distances. Finally, it assigns a score to each reference haplotype that quantifies the likelihood that the original alleles are emitted from the respective reference haplotype. After all haplotypes are scored for all variants, Viterbi decoding is used to trace back the highest scoring decoded allele sequence for the proxy haplotype. Given that each variant is independently hashed, this procedure will find the haplotypes with largest joint probabilistic score that combines the haplotype frequency concordance between the proxy panel and the reference panel, and the recombination rates of consecutive SNPs.

Of note, the computational resources for executing this HMM is at least as large as a Viterbi-based imputation HMM. Furthermore, the state reduction techniques[68] are not directly applicable to the proxy-decoding HMM because the state-switching probabilities rely on both the variant location and the k-mer concordance. Therefore, unlike Homer's attack and LRT attacks, which can be easily performed by honest-but-curious entities, proxy-decoding attacks require large computational resources with a malicious intent.

To test accuracy of the re-identification using decoded proxy panels, we divided the 661 subjects in AFR population of the 1000 Genomes Project into three panels. The first panel (100 subjects) is used to generate a proxy panel. The second panel (461 subjects) is used as the reference panel for proxy-decoding and LRT attack. The third panel (100 subjects) is held out as a control for the baseline LRT attack on non-matching individuals (control panel). We focused on the 87,960 variants on chromosome 1. Since the variants used by the proxy panels are obfuscated, we divided the variants into two categories by assigning every other variant to the proxy panel and the remaining variants to the reference panel such that the panels contained 43,980 non-overlapping variants. This way, we assume that the attacker approximately guesses the positions of the proxy panel variants. To decrease computational requirements, we focused on the variants between 10-100 million base pairs, which results in 19,379 variants. We generated the proxy panel for the first panel and decoded the proxy-haplotypes using the reference panel. We next evaluated the accuracy of the decoded alleles (for proxy panel variants) by calculating the concordance of the known original panel alleles and the decoded alleles. As a baseline control, we shuffled the decoded panel subjects and calculated the average allelic difference between shuffled decoded panel and the first panel. Overall, the proxy-decoding procedure does provide improvement of around 7% over control (30% vs 23%) (Supplemental Fig. S5a) when small hashing window length is used (hashing with 7-mers). As the hashing window length is increased to 17-mers, the accuracy of decoded proxy is 3% lower than control (30% vs 27%) (Supplemental Fig. S5a).

LRT Attack using Decoded Proxy Panel. We next used the proxy decoded alleles of the proxy panel in LRT-based re-identification attack. For this, we used the decoded proxy panel as a pool in which the adversary tries to re-identify an individual whose genome is available. The reference panel (461 subjects) used for decoding is again used as the reference panel for the LRT-based attack. For each original panel subject, we calculate the LRT statistic. As a control to individuals in the pool (original panel), we calculated the LRT statistic for the subjects in the held-out control panel (100 subjects), who did not participate in the pool. When the LRT statistics are compared, we did not find a clear separation between the distribution of LRT scores for the original panel subjects and the control panel subjects (Supplemental Fig. 5b). When we applied LRT on the cleartext panels (no proxy panels), the test statistic clearly distinguishes the original and control panels for re-identifiability (Supplemental Fig. S5c). We also calculated LRT test on the sampled original panel without hashing, i.e., usage of mosaic haplotypes only and observed that original panel can be easily distinguished from control panel (Supplemental Fig. S5d). However, the test may be miscalibrated (LRT statistic is below 0 for approximately half of the original panel subjects). This result demonstrates the importance of using the allele hashing while sharing haplotype datasets. We also observed that when the population of the reference panel in LRT attack does not match to the ancestry of the proxy panel, it is challenging to calibrate LRT attack to re-identify individuals. This is concordant with previous studies that highlighted the importance of matching ancestry of the reference panel to the target individual and the pool that is being tested[38,39]. Unless the adversary has the knowledge of exact ancestral groups of the individuals in proxy panel, we expect that the re-identification will be more challenging than our approach. Notably, our previous results show that a PCA analysis does not immediately reveal ancestral status of subjects in the proxy panel (Supplemental Fig. S4j). This renders it more challenging for an adversary to build a reference panel for re-identification purposes.

Emission Probability and Imputation-based Attacks on RESHAPE'd panels

RESHAPE uses a sampling approach to generate mosaicize'd panels from phased panels. The sampling relies on one parameter of RESHAPE, which is the number of generations. The number of generations tunes the number of recombinations that are used in samplings. It is expected that the higher number of generations should provide strong privacy protections. In the imputation accuracy tests, we used number of generations set to 128, which was deemed to be a strong protection parameter in the RESHAPE study.

Emission Probabilities of RESHAPE'd reference panels can re-identify individual participation.

Similar to the previous decoding attack scenario on ProxyTyper, we asked whether the participation of individuals can be identified when we are given a RESHAPE'd panel. For this, we implemented a forward algorithm to calculate the emission probability of a phased genome. This is very simple algorithm that calculates the total probability over all of the Markov chain paths using Li-Stephens recombination probabilities.

To test this simple emission probability test, we randomly selected 5000 subjects from the Haplotype Reference Consortium panel, which represent the reference panel that will be protected via RESHAPE. We also selected 200 subjects from the reference panel that we will use as the positive set from within the protected reference panel. We next selected 200 subjects as the holdout (negative) set from HRC panel, which does not overlap with the 5,000 subject reference panel.

We selected 20,000 random variants from the region 20:10,000,000-20,000,000 region. We first ran RESHAPE to protect the 5,000 subject reference panel using number of generations set to 128 generations. We calculated the emission probabilities of every subject in positive and negative panels (400 subjects in total) using the RESHAPE'd panel as the reference panel, which is shown in Supplemental Figure 6a. We observe a clear separation between the positive and negative panels.

Simple Statistics may Inadvertently be used to re-identify individual Participation.

The emission probability attack we demonstrated above may be deemed too complex for an honest-but-curious entity.

We therefore asked if there are simpler statistics that can inadvertently leak information to a curious entity. We first emphasize that the emission probabilities that we calculated above ae based on the forward algorithm, which is very efficiently calculated within imputation tools such as BEAGLE. In fact, BEAGLE performs forward-backward calculations much more efficiently than our implementation. These forward-backward variables are used to calculate the posterior probabilities of imputed alleles. Using an educated guess, we reasoned that the emission probabilities should correlate with the emission probabilities from a simple BEAGLE run.

To test this expectation, we used 10,000 subjects and 150,000 variants and ran RESHAPE using the number of generations in the set {128, 256, 512}. From within 10,000 subjects, we randomly selected 1,000 subjects as the positive subjects. We similarly selected 1,000 non-overlapping subjects as the holdout negative subject set.

For each of the positive/negative panel, we set randomly selected 75,000 variants aside as typed variants and ran BEAGLE to impute the remaining 75,000 variant variants where the positive and negative sets were used as the panels with typed variants and the RESHAPE'd panel with 10,000 subjects (with 150,000 variants) as the reference panel.

We finally wrote an awk script to calculate the average of the total imputed allele frequency (which is a floating-point value between 0 and 2) for each subject, which are shown as box plots in Supplemental Figure S6b. It can be seen that there is strong separation between the average imputed allele frequencies of the variants of positive and negative subjects. As expected, the positive subjects have substantially higher imputed allele frequencies compared to the negative set. This is expected since subjects in the positive set have their genome integrated into the RESHAPE'd panel. Thus, when we use the RESHAPE'd panel as a reference, the imputed probabilities are much higher for the subjects who were in the RESHAPE'd panel compared to the subjects who were not included in the RESHAPE'd panel (i.e., negative panel.) Interestingly, we observe a strong separation for 256 generations (Supplemental Figure S6c) and separation is still observed for 512 generations (Supplemental Figure S6d). Overall, these results indicate that releasing the coordinates and alleles of variants may inadvertently lead to an attack surface that can be used to leak information that may be deemed privacy breaching.

We do not consider any further linking attacks since these would not accomplish any reasonable purpose. We, however, do believe plaintext release of variant coordinates and alleles can be problematic under numerous known genealogy data probing attacks for re-identifying individuals.

References

1. Bonomi L, Huang Y, Ohno-Machado L. Privacy challenges and research opportunities for genomic data sharing. Nat Genet. 2020;52: 646–654.

2. Yousefi S, Abbassi-Daloii T, Kraaijenbrink T, Vermaat M, Mei H, van 't Hof P, et al. A SNP panel for identification of DNA and RNA specimens. BMC Genomics. 2018;19. doi:10.1186/s12864-018-4482-7

3. Hafiza MK, Sajjad A, Nasir S, Qazi LA, Muhammad A, Mohammad AT. Exoneration of primary suspect after false confession with the help of forensic DNA analysis. Forensic Genom. 2022;2: 17–20.

4. Seydel C. You can run, but your DNA can't hide. Forensic Genom. 2022;2: 97–102.

5. Niemiec E, Howard HC. Ethical issues in consumer genome sequencing: Use of consumers' samples and data. Appl Transl Genom. 2016;8: 23–30.

6. Pulivarti R. Cybersecurity of Genomic Data. Gaithersburg, MD: National Institute of Standards and Technology; 2023. doi:10.6028/nist.ir.8432.ipd

7. Sherburn IA, Finlay K, Best S. How does the genomic naive public perceive whole genomic testing for health purposes? A scoping review. Eur J Hum Genet. 2023;31: 35–47.

8. Jamal L, Sapp JC, Lewis K, Yanes T, Facio FM, Biesecker LG, et al. Research participants' attitudes towards the confidentiality of genomic sequence information. Eur J Hum Genet. 2014;22: 964–968.

9. After Havasupai litigation, Native Americans wary of genetic research. Am J Med Genet A. 2010;152A: fmix.

10. Garrison NA. Genomic justice for native Americans: Impact of the Havasupai case on genetic research. Sci Technol Human Values. 2013;38: 201–223.

11. Powell K. The broken promise that undermines human genome research. Nature. 2021;590: 198–201.

12. Walking the tightrope between data sharing and data protection. Nat Med. 2022;28: 873.

13. Budin-Ljøsne I, Isaeva J, Knoppers BM, Tassé AM, Shen H-Y, McCarthy MI, et al. Data sharing in large research consortia: experiences and recommendations from ENGAGE. Eur J Hum Genet. 2014;22: 317–321.

14. Stobbe MD, Gonzalez-Perez A, Lopez-Bigas N, Gut IG. Ten simple rules for a successful international consortium in big data omics. PLoS Comput Biol. 2022;18: e1010546.

15. Cohen IG, Mello MM. HIPAA and Protecting Health Information in the 21st Century. JAMA. 2018;320: 231–232.

16. Greenbaum D, Sboner A, Mu XJ, Gerstein M. Genomics and privacy: Implications of the new reality of closed data for the field. PLoS Computational Biology. 2011. doi:10.1371/journal.pcbi.1002278

17. Hubaux J-P, Katzenbeisser S, Malin B. Genomic data privacy and security: Where we stand and where we are heading. IEEE Secur Priv. 2017;15: 10–12.

18. Wan Z, Hazel JW, Clayton EW, Vorobeychik Y, Kantarcioglu M, Malin BA. Sociotechnical safeguards for genomic data privacy. Nat Rev Genet. 2022;23: 429–445.

19. Erlich Y, Narayanan A. Routes for breaching and protecting genetic privacy. Nat Rev Genet. 2014;15: 409–421.

20. Erlich Y, Williams JB, Glazer D, Yocum K, Farahany N, Olson M, et al. Redefining genomic privacy: trust and empowerment. PLoS Biol. 2014;12: e1001983.

21. Shabani M, Marelli L. Re-identifiability of genomic data and the GDPR: Assessing the re-identifiability of genomic data in light of the EU General Data Protection Regulation. EMBO Rep. 2019;20: e48316.

22. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. Science. 2013;339: 321–324.

23. Lin Z, Owen AB, Altman RB. Genetics. Genomic research and human subject privacy. Science. 2004;305: 183.

24. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet. 2008;4: e1000167.

25. Sankararaman S, Obozinski G, Jordan MI, Halperin E. Genomic privacy and limits of individual detection in a pool. Nat Genet. 2009;41: 965–967.

26. Visscher PM, Hill WG. The limits of individual identification from sample allele frequencies: theory and statistical analysis. PLoS Genet. 2009;5: e1000628.

27. Im HK, Gamazon ER, Nicolae DL, Cox NJ. On sharing quantitative trait GWAS results in an era of multiple-omics data and the limits of genomic privacy. Am J Hum Genet. 2012;90: 591–598.

28. Harmanci A, Gerstein M. Quantification of private information leakage from phenotype-genotype data: linking attacks. Nat Methods. 2016;13: 251–256.

29. Harmanci A, Gerstein M. Analysis of sensitive information leakage in functional genomics signal profiles through genomic deletions. Nat Commun. 2018;9. doi:10.1038/s41467-018-04875-5

30. Branum R, Wolf SM. International policies on sharing genomic research results with relatives: Approaches to balancing privacy with access. J Law Med Ethics. 2015;43: 576–593.

31. Telenti A, Ayday E, Hubaux JP. On genomics, kin, and privacy. F1000Res. 2014. doi:10.12688/f1000research.3817.1

32. Bu D, Wang X, Tang H. Haplotype-based membership inference from summary genomic data. Bioinformatics. 2021;37: i161–i168.

33. Jacobs KB, Yeager M, Wacholder S, Craig D, Kraft P, Hunter DJ, et al. A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies. Nat Genet. 2009;41: 1253–1257.

34. Fiume M, Cupak M, Keenan S, Rambla J, de la Torre S, Dyke SOM, et al. Federated discovery and sharing of genomic data using Beacons. Nat Biotechnol. 2019;37: 220–224.

35. Shringarpure SS, Bustamante CD. Privacy Risks from Genomic Data-Sharing Beacons. Am J Hum Genet. 2015;97: 631–646.

36. Ayoz K, Ayday E, Cicek AE. Genome reconstruction attacks against genomic data-sharing beacons. Proc Priv Enhancing Technol. 2021;2021: 28–48.

37. Thenen NV, Ayday E, Cicek AE. Re-Identification of Individuals in Genomic Data-Sharing Beacons via Allele Inference. Bioinformatics. 2018. doi:10.1101/200147

38. Egeland T, Fonneløp AE, Berg PR, Kent M, Lien S. Complex mixtures: a critical examination of a paper by Homer et al. Forensic Sci Int Genet. 2012;6: 64–69.

39. Sampson J, Zhao H. Identifying individuals in a complex mixture of DNA with unknown ancestry. Stat Appl Genet Mol Biol. 2009;8: Article 37.

40. Dwork C, Roth A. The algorithmic foundations of differential privacy. Found Trends Theor Comput Sci. 2013;9: 211–407.

41. Dwork C. Differential privacy: A cryptographic approach to private data analysis. Privacy, Big Data, and the Public Good. New York: Cambridge University Press; 2014. pp. 296–322.

42. Kim M, Harmanci AO, Bossuat J-P, Carpov S, Cheon JH, Chillotti I, et al. Ultrafast homomorphic encryption models enable secure outsourcing of genotype imputation. Cell Systems. 2021;12: 1108-1120.e4.

43. Yang M, Zhang C, Wang X, Liu X, Li S, Huang J, et al. TrustGWAS: A full-process workflow for encrypted GWAS using multi-key homomorphic encryption and pseudorandom number perturbation. Cell Syst. 2022;13: 752-767.e6.

44. Froelicher D, Troncoso-Pastoriza JR, Raisaro JL, Cuendet MA, Sousa JS, Cho H, et al. Truly privacy-preserving federated analytics for precision medicine with multiparty homomorphic encryption. Nat Commun. 2021;12: 5910.

45. TrustGWAS: A full-process workflow for encrypted genome-wide association studies using multi-key homomorphic encryption and pseudo-random number perturbation. Github; Available: https://github.com/melobio/TrustGWAS

46. Cho H, Wu DJ, Berger B. Secure genome-wide association analysis using multiparty computation. Nat Biotechnol. 2018;36: 547–551.

47. Blatt M, Gusev A, Polyakov Y, Rohloff K, Vaikuntanathan V. Optimized homomorphic encryption solution for secure genome-wide association studies. BMC Med Genomics. 2020;13: 83.

48. Shimizu K, Nuida K, Rätsch G. Efficient privacy-preserving string search and an application in genomics. Bioinformatics. 2016;32: 1652–1661.

49. Nakagawa Y, Ohata S, Shimizu K. Efficient privacy-preserving variable-length substring match for genome sequence. Algorithms Mol Biol. 2022;17: 9.

50. Popic V, Batzoglou S. A hybrid cloud read aligner based on MinHash and kmer voting that preserves privacy. Nat Commun. 2017. doi:10.1038/ncomms15311

51. Gentry C. A FULLY HOMOMORPHIC ENCRYPTION SCHEME. PhD Thesis. 2009; 1–209.

52. Kim M, Lauter K. Private genome analysis through homomorphic encryption. BMC Med Inform Decis Mak. 2015;15 Suppl 5: S3.

53. Dowlin N, Gilad-Bachrach R, Laine K, Lauter K, Naehrig M, Wernsing J. Manual for using homomorphic encryption for bioinformatics. Proc IEEE Inst Electr Electron Eng. 2017;105: 1-16;

54. Orlandi C. Is multiparty computation any good in practice? ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. 2011. doi:10.1109/ICASSP.2011.5947691

55. Zhao C, Zhao S, Zhao M, Chen Z, Gao C-Z, Li H, et al. Secure Multi-Party Computation: Theory, practice and applications. Inf Sci (Ny). 2019;476: 357–372.

56. Wang W, Chen G, Pan X, Zhang Y, Wang XF, Bindschaedler V, et al. Leaky cauldron on the dark land: Understanding memory side-channel hazards in SGX. Proceedings of the ACM Conference on Computer and Communications Security. New York, NY, USA: Association for Computing Machinery; 2017. pp. 2421–2434.

57. Nilsson A, Bideh PN, Brorsson J. A survey of published attacks on Intel SGX. arXiv. 2020. Available: http://arxiv.org/abs/2006.13598

58. Bernier A, Liu H, Knoppers BM. Computational tools for genomic data de-identification: facilitating data protection law compliance. Nat Commun. 2021;12: 6949.

59. Heeney C, Hawkins N, de Vries J, Boddington P, Kaye J. Assessing the privacy risks of data sharing in genomics. Public Health Genomics. 2011;14: 17–25.

60. Gonzales A, Guruswamy G, Smith SR. Synthetic data in health care: A narrative review. PLOS Digit Health. 2023;2: e0000082.

61. Cavinato T, Rubinacci S, Malaspinas A-S, Delaneau O. A resampling-based approach to share reference panels. bioRxiv. 2023. p. 2023.04.07.535812. doi:10.1101/2023.04.07.535812

62. Yelmen B, Decelle A, Ongaro L, Marnetto D, Tallec C, Montinaro F, et al. Creating artificial human genomes using generative neural networks. PLoS Genet. 2021;17: e1009303.

63. Wohns AW, Wong Y, Jeffery B, Akbari A, Mallick S, Pinhasi R, et al. A unified genealogy of modern and ancient genomes. Science. 2022;375: eabi8264.

64. Anderson-Trocmé L, Nelson D, Zabad S, Diaz-Papkovich A, Kryukov I, Baya N, et al. On the genes, genealogies, and geographies of Quebec. Science. 2023;380: 849–855.

65. Raisaro JL, Tramèr F, Ji Z, Bu D, Zhao Y, Carey K, et al. Addressing Beacon re-identification attacks: quantification and mitigation of privacy risks. J Am Med Inform Assoc. 2017;24: 799–805.

66. Li N, Stephens M. Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. Genetics. 2003;165: 2213–2233.

67. Ycart B. Letter counting: a stem cell for Cryptology, Quantitative Linguistics, and Statistics. arXiv [math.HO]. 2012. Available: http://arxiv.org/abs/1211.6847

68. Wang S, Kim M, Jiang X, Harmanci AO. Evaluation of vicinity-based hidden Markov models for genotype imputation. BMC Bioinformatics. 2022;23. doi:10.1186/s12859-022-04896-4