GENOFA: QUANTIFYING PRIVACY LEAKAGE OF GENOMIC LANGUAGE MODELS UNDER DEPENDENT DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Genomic language models (gLMs) enable large-scale training on personal genomic variants and advance personalized medicine. However, they also introduce unique privacy risks. Human genomes are inherently correlated across familial relationships due to Mendelian inheritance. Such correlation breaks the assumption of independence required by most differentially private algorithms, preventing straightforward adoption in the context of genomic privacy. We leverage Mendelian inheritance to create a family-aware genotype reconstruction attack, called GENOFA. gLMs can *impute* on non-training data points, but this should not count as privacy leakage as it occurs for seen and unseen data alike; we propose imputation-corrected reconstruction risk, a novel metric to remove the model's imputation impact from privacy accounting. Our results demonstrate that familial correlations can amplify individual privacy leakage, motivating the need for family-level privacy protections. We therefore propose to treat the *family* as an atomic unit for privacy protection and apply client-level differential privacy to bound the privacy leakage of the entire family. We demonstrate that with a privacy budget of 6.0, client-DP improves the model utility by 22.9%, reducing the RMSE from 0.35 (group-DP) to 0.27. Together, these contributions provide a unified framework that identify, quantify, and mitigate privacy leakage in gLMs despite the inclusion of familially related individuals.

1 INTRODUCTION

Genomics language models (gLMs) leverage DNA sequences as input to learn contextual representations of the genome through self-supervised objectives. Recent studies (Ghosal et al., 2025; Dalla-Torre et al., 2025; Fishman et al., 2025) demonstrate the benefits of training gLMs on personal *variant genomes* to predict personal traits. Unlike a generalized, species-wide reference genome, variant genomes represent an individual's unique DNA sequence (Aganezov et al., 2022).

Genotypes, the specific genetic material at those variant locations can influence disease risk, treatment response, and other clinically relevant traits, making variant genomes a promising foundation for personalized medicine. However, releasing or modelling this sensitive data raises severe privacy risks, including de-anonymization (Thomas et al., 2024; Harmanci & Gerstein, 2016), membership inference (MIA) (Homer et al., 2008; von Thenen et al., 2019), and reconstruction attacks (Fredrikson et al., 2014; Paige et al., 2020).

Genomic studies often include family-related individuals, yet the resulting privacy implications remain understudied. We characterize the risks of training gLMs on family-correlated data. We introduce GENOFA, a novel family-aware genotype reconstruction attack. Our attack explicitly exploits inherent genomic correlations between relatives via Mendelian Inheritance to infer an unknown genotype of a target individual. Compared to a naive baseline, GENOFA increases training data reconstruction accuracy from 62% to 75% (Table 1). However, we cannot attribute the higher attack success solely to the inclusion of the target in the training set, because we also observe an increase in reconstruction accuracy for unseen validation data (i.e., imputation) from 36% to 40%

To eliminate the impact of the model's imputation capability in privacy accounting, we propose a novel metric, *imputation-corrected reconstruction risk*. Consider Table 1, the imputation-corrected privacy leakage for GENOFA is 35% (75% reconstruction success minus 40% imputation success) which is 9% more than the privacy leakage of the naive attack (26%). The actual metric is more

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

nuanced but carries the same intention: instead of using accuracy, we leverage the reconstruction probability to capture the attack success rate difference on training and non-training data.

With the proposed privacy quantification metric, we show elevated privacy leakage due to the inclusion of family members in training. Standard privacy-preserving ML training algorithms, such as differentially private stochastic gradient descent (Abadi et al., 2016) protect privacy at the individual record level; such algorithms do not protect against the additional leakage caused by correlation. To defend against such correlation-induced attacks, we propose to take the family as the unit of privacy and leverage client-level differentially private algorithms as a family-aware privacy-preserving training algorithm. We also propose an empirical audit procedure, *group-LiRA*, that evaluates family-level privacy, given that existing privacy audit methods (Shokri et al., 2017; Carlini et al., 2022) primarily target individual records rather than families. When finetuning the gLMs on the 1000G (Byrska-Bishop et al., 2022) database at a privacy budget of 6.0, the family-aware privacy-preserving training algorithm reduces family-level privacy leakage by over 40% while preserving strong model utility (against validation RMSE of 0.27).

Table 1: Reconstruction accuracy (%) of naive and GENOFA attacks on training (family members and target are in training) vs. validation (family members and target are not in training; family members are used as auxiliary information).

Split	Naive	GENOFA
Train	61.76	75.00
Val (Imputation)	36.15	40.00

2 BACKGROUND AND PRELIMINARIES

In this work, we focus on privacy attacks and privacy-preserving mechanisms for machine learning in genomics. We begin by introducing the structure and terminology of genomic data.

2.1 PERSONAL GENOMIC VARIANTS DATASET STRUCTURE

Terminology. Humans share roughly 99.9% of their DNA with the reference genome Aganezov et al. (2022); the remaining 0.1% consists of genetic variants. At each variant location, an individual carries specific *alleles*, i.e., variations in the nucleotides A, C, G, and T. Because each parent contributes one chromosome set (a *haplotype*), individuals possess two paired haplotypes. The specific pair of alleles at a given variant location constitutes the *genotype*. These genetics influence the *phenotype*, which is any observable physical or molecular trait, such as height, disease status, or gene expression.

Tabular formulation of genomic data. Since most variants occur at publicly documented loci, genomes aligned to a standard reference assembly can be mapped into a table, as in Table 2. Each individual is a single row of N attributes, with values representing genotypes and study phenotypes.

Personal genomic sequences are constructed by modifying the reference sequence with an individual's variant alleles. Depending on the studied variants and the context window of the gLM, an individual's tuple may be fragmented across M localized windows, a fundamental difference from conventional tabular formats where records remain indivisible. Without loss of generality, we focus our analysis on supervised learning (mapping sequences to phenotypic labels), though this naturally extends to semi-supervised settings by omitting labels. Further data curation details are in Appendix Section A.11.

Genomes of family members are correlated via Mendelian inheritance. Because a child's genotype is derived directly from their parents, familial records are statistically dependent regardless of whether they appear in the same study. For example, consider two genomic variant locations where a father's genotype is `A|a, b|B` (where "|" separates alleles by parent of origin) and the mother's is `A|A, B|B`. Assuming a crossover (i.e., the exchange of alleles between paired chromosomes) rate of r and no *de novo* mutations, the mother deterministically transmits `AB`. The father transmits `Ab`, `AB`, `aB`, or `ab` with probabilities $\frac{1}{2}(1-r)$, $\frac{1}{2}r$, $\frac{1}{2}(1-r)$, and $\frac{1}{2}r$, respectively. Placing the paternally inherited alleles before the "|" yields: $\Pr[\text{kid} = g \mid \text{d} = \texttt{A|a,b|B}, \text{m} = \texttt{A|A,B|B}] = \frac{1}{2}(1-r)$ if $g = \texttt{A|A,b|B}$ or $\texttt{a|A,B|B}$; otherwise $\frac{1}{2}r$ (for $g = \texttt{A|A,B|B}$ or $\texttt{a|A,b|B}$).

If we know the kid's genotype is `A|A`, [unknown], the probability of the kid's unknown genotype given knowledge of their parents' genotypes and of the kid's other known genotypes is captured with:

$$\Pr[\text{kid} = g \mid \text{d} = \texttt{A|a,b|B}, \text{m} = \texttt{A|A,B|B}, \text{kid} = \texttt{A|A}, [\text{unknown}]] = \begin{cases} (1-r) & g = \texttt{A|A,b|B} \\ r & g = \texttt{A|A,B|B} \end{cases} \tag{1}$$

Table 2: Genotype *Phased* Data for M individuals across N genomic variant locations and the phenotypical label. A reference allele=0; 1 and above represent alternative alleles in the order. The variant locations and phenotypical label are attributes. Each variant location contains one genotype value (i.e., one allele from each parent concatenated by "|").

Sample ID H1 | H2	**Variant Loc 1** ref=A, alt=G	**Variant Loc 2** ref=C, alt_1=T alt_2=G	**Variant Loc 3** ref=C, alt=$\emptyset$	...	**Variant Loc N** ref=G, alt=CG	**Phenotype** (label) e.g., cell expression
Person 1	0 | 1	0 | 1	1 | 1	...	0 | 1	0.5
⋮	⋮	⋮	⋮	⋱	⋮	...
Person M	0 | 0	0 | 0	1 | 0	...	1 | 1	0.0

In this work we characterize how the Mendelian inheritance increases the privacy risk of individuals from the same family. But first we need to formally define our privacy notion.

2.2 Differential Privacy

Differential Privacy (DP) (Dwork & Roth, 2014; Dwork, 2011; Dwork et al., 2006) is the standard notion to reason about the privacy guarantees of algorithms that operate on sensitive datasets.

Definition 1. (ϵ, δ)**-Differential Privacy:** A randomized mechanism $M : \mathscr{X} \to \mathscr{R}$ that maps from the input space $\mathscr{X}$ to output space $\mathscr{R}$ satisfies (ϵ, δ)-DP if for any adjacent datasets $X, X' \in \mathscr{X}$ (X and X' differ by one data record) and $S \subset \mathscr{R}$: $\Pr[M(X) \in S] \leq e^{\epsilon} \Pr[M(X') \in S] + \delta$.

The parameter ϵ is usually referred to as the privacy budget. In the definition, adjacent datasets X and X' differ by a single data record. This formulation is known as record-DP. This adjacency defines the privacy granularity, i.e., the unit of privacy protection. Traditionally, each individual contributes a single record, making the record the standard unit of privacy in genomics (Uhler et al., 2013; Lange, 2024). However, for gLM training, one person's genome may be segmented into multiple data points, hence the person's privacy leakage is the total leakage of all data points contributed by the person. The standard DP training algorithm for ML is the DP-Stochastic Gradient Descent (DP-SGD) (Abadi et al., 2016), which provides record-level privacy protection. We reproduce the algorithm in Appendix A.3.

2.3 Related Work

While privacy leakage is extensively studied in general machine learning, including membership inference attacks (MIA, which infer if a data record was used in model training) (Carlini et al., 2021; Shokri et al., 2017), attribute inference (which reconstructs unknown attribute values) (Fredrikson et al., 2014; Yeom et al., 2018), and privacy guarantee auditing (Steinke et al., 2023), the privacy risks of gLMs remain largely under-explored. Although recent studies demonstrate that genomic training segments are vulnerable to MIA and data extraction (Belfiore et al., 2025; Nemecek et al., 2026), they omit family structures, thereby overlooking how relatedness correlations amplify leakage when analyzing inherited traits. Furthermore, these works implicitly assume record-level granularity (e.g., genomic sub-segments), requiring additional steps to compute total per-person risk.

To bridge this gap, our framework treats the family as the fundamental unit of privacy granularity, mitigating correlation-induced leakage when relatives are present in the training data while naturally reducing to standard person-level privacy for single-member families. In the closest prior work, Almadhoun et al. (2020) reconstructed target genotypes using aggregate familial allele counts, but limited their analysis to summation queries. We extend this evaluation to gLMs, which are inherently capable of complex data imputation, and introduce a novel metric to adjust for these heightened imputation risks. More related work can be found in Appendix A.12.

3 Method: Proposed Family-Aware Attack (GenoFA) and Family-Aware Privacy Protection

In this section, we first introduce the threat model with a novel attribute inference attack for genotype reconstruction in the style of Yeom et al. (2018). We then outline the baseline attack, which does not leverage family structure. Subsequently, we introduce the proposed family-aware attack, GenoFA, which leverages Mendel's laws and observed relatives to reconstruct the geno-

type of the target. Then, we propose a novel metric to quantify such privacy leakage, namely imputation-corrected reconstruction risk. Lastly, we propose to treat the entire family as an atomic unit and to train the models with family-level privacy protection.

3.1 THREAT MODEL

For data distribution $\mathscr{D}$, person i's data $(\boldsymbol{x}^i, y^i) \in \mathscr{D}$, $\boldsymbol{x}^i = x^i_{1:N}$ is the N input genotypes to the model and y^i is the output phenotypical label. To observe the worst-case privacy leakage, we assume a strong adversary that has *maximum amount of knowledge* about the target individual, the released model, and the population distribution. That is, we assume there is one single unknown *target* genotype x^i_t that the adversary tries to reconstruct. The adversary knows $\mathcal{G}_t$, which is the set of all possible values for x^i_t; the adversary also knows all other genotypes within the genomic range of interest, $x^i_{-t} = x^i_{1:t-1,t+1:N}$. Also, the adversary has access to the released regression model f such that $y^i = f(\boldsymbol{x}^i) + \xi^i$ where $\xi^i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ is the error distribution (Gaussian). The adversary knows population information about the target genotype (i.e., $\Pr[x_t]$ from the data distribution). Let $x^{i*}_t \in \mathcal{G}_t$ be the genotype reconstructed by the adversary. Let $g_{\text{true}} \in \mathcal{G}_t$ be the ground truth value of x^i_t.

For the GENOFA (family-aware) attacks, we assume the adversary additionally knows Ψ of an individual i, which is the family structure (i.e., who the parents and the kids are) and the known genotypes and labels of the for relatives of i, i.e., $\Psi := \{x^j_{-t}, y^j\}, \ \ \forall j \in i$'s family. For the attack not using the family structure, $\Psi = \emptyset$.

Definition 2 (Attribute Inference Attack Experiment (Yeom et al., 2018))**.** Let $\mathcal{A}$ be the adversary (attack method).

1. Sample a training dataset D_{train} from $\mathscr{D}$ and train f using D_{train}.
2. Choose $b \leftarrow \{0, 1\}$ uniformly at random.
3. Sample a target person $(\boldsymbol{x}^i, y^i)$ from $\mathscr{D}$ if $b = 1$, else sample from D_{train}.
4. The attack $\mathcal{A}$ takes all known information, $x^i_{-t}, y^i, f, \mathscr{D}$, and Ψ, and infer the value of the unknown genotype x^i_t: $x^{i*}_t := \mathcal{A}(x^i_{-t}, y^i, f, \mathscr{D}, \Psi)$. The experiment returns $\mathbb{1}\left[x^{i*}_t = g_{\text{true}}\right]$, where $\mathbb{1}[\circ]$ is the indicator function.

The actual algorithms of the attack $\mathcal{A}$, baseline (family-unaware) and the proposed GENOFA (family-aware) are defined in Sections 3.2 and 3.3. To make sure the target person's family also in D_{train} when $b = 0$ are not in D_{train} when $b = 1$, families are sampled as atomic units to construct D_{train}.

3.2 BASELINE NAIVE ATTACK (FREDRIKSON ET AL., 2014)

For $g \in \mathcal{G}_t$, we define the model-prediction likelihood $L_i(g) := \Pr[y^i \mid x^i_t = g, x^i_{-t}] = \mathcal{N}\big(y^i;\, f(g, x^i_{-t}),\, \sigma^2\big)$, and the prior $\pi_i(g) := \Pr[x^i_t = g \mid x^i_{-t}]$. The naive per-person reconstruction likelihood is $\Pr_{\text{naive}}[x^i_t = g \mid y^i, x^i_{-t}] = \frac{L_i(g)\,\pi_i(g)}{\sum_{g' \in \mathcal{G}_i} L_i(g')\,\pi_i(g')}$, and the ground truth value for x^i_t is g_{true}. The naive attack returns

$$\mathcal{A}_{\text{naive}}(x^i_{-t}, y^i, f, \mathscr{D}, \Psi = \emptyset) = \arg\max_g \Pr_{\text{naive}}[x^i_t = g \mid y^i, x^i_{-t}]. \tag{2}$$

We assume independence among input attributes following Fredrikson et al. (2014): $\pi_i(g) \approx \Pr[x^i_t = g]$, estimated from $\mathscr{D}$. Since this prior is shared by both naive and GenoFA attacks, any performance gap reflects solely the knowledge gained from Mendelian transmission $T(k \mid m, d)$. By omitting linkage disequilibrium (LD), defined as the statistical correlation between genetic variants across a population, we establish a tractable baseline that isolates familial data leakage from background population structure. We discuss the implications of this independence assumption, along with strategies for integrating LD into the model, in Section 5.

3.3 GENOFA ATTACK: TARGETING THE FAMILY-DEPENDENCY

Similar to the naive attack, GENOFA reconstructs a target genotype value of individual i using their known genotypes (x^i_{-t}), label (y^i), and the model weights and error distribution(f and σ). However, GENOFA further leverages $\Psi := \{x^j_{-t}, y^j\}, \ \ \forall j \in i$'s family, the familial information from the target's family. Here, we demonstrate GENOFA using a family of three.

Example: Let $\boldsymbol{x}^1$, $\boldsymbol{x}^2$, $\boldsymbol{x}^3$ be the input genotypes of the kid, mother, and father respectively, $\boldsymbol{y} = (y^1, y^2, y^3)$ be their labels. The information flows as $\boldsymbol{x}^2 \rightarrow \boldsymbol{x}^1 \leftarrow \boldsymbol{x}^3$ (the parents are in-

dependent; the kid depends on both). Let $\mathbf{x}_{-t} = (x^1_{-t}, x^2_{-t}, x^3_{-t})$ be their known genotypes, whereas x^1_t, x^2_t, x^3_t are the unknown genotype to be reconstructed. As the example in Eq 1, we define Mendelian transmission $T(k \mid m, d) := \Pr[x^1_t = k \mid x^2_t = m, x^3_t = d, \mathbf{x}_{-t}]$ to model the probability of a kid's unknown genotype given the parental genomes and the known genotypes of the kid.

The probability for the unknown genotypes for the family given their known genotypes is $\Pr[x^1_t = k, x^2_t = m, x^3_t = d \mid \boldsymbol{y}, \mathbf{x}_{-t}] \propto \Pr[\boldsymbol{y} \mid k, m, d, \mathbf{x}_{-t}] \Pr[k, m, d \mid \mathbf{x}_{-t}] = L_1(k)L_2(m)L_3(d) \underbrace{T(k \mid m, d)}_{\text{Mendelian}} \underbrace{\pi_2(m)\,\pi_3(d)}_{\text{parental priors}}$, GENOFA reconstruction likelihood is

$\Pr_{\text{GENOFA}}[x^i_t = g \mid y^i, x^i_{-t}, \Psi] = \Pr[x^i_t = g \mid \boldsymbol{y}, \mathbf{x}_{-t}] = \sum_{x^j_t, j \neq i} \Pr[x^1_t, x^2_t, x^3_t \mid \boldsymbol{y}, \mathbf{x}_{-t}]$, where $\Psi := \{x^j_{-t}, y^j\}$, $\forall j \in i$'s family. Let the ground truth value for x^i_t be g_{true}, GENOFA returns

$$\mathcal{A}_{\text{GENOFA}}\left(x^i_{-t}, y^i, f, \mathscr{D}, \Psi = \{x^j_{-t}, y^j\}\right) = \arg\max_g \Pr_{\text{GENOFA}}[x^i_t = g \mid y^i, x^i_{-t}, \Psi]. \tag{3}$$

Detailed derivation and families of two parents and multiple kids are provided in Appendix A.2.

3.4 QUANTIFY THE PER-PERSON PRIVACY LEAKAGE

Typically, the privacy leakage of the attribute inference attack is quantified using the reconstruction accuracy defined in Definition 2, as motivated in Section 1. Using such hard outputs loses information about the reconstruction likelihood. Hence, following Carlini et al. (2022), we use a soft output to quantify the reconstruction score from Eq. 2 and Eq. 3. Recall $\Psi = \emptyset$ when $\mathcal{A}$ =naive; $\Psi = \{x^j_{-t}, y^j\}$, $\forall j \in$ the family of target i when $\mathcal{A}$ =GENOFA.

Definition 3 (Reconstruction Score $\mathcal{A}'$)**.** With the normalized logits $\phi(p) := \log \frac{p}{1-p}$, the reconstruction score is $\mathcal{A}'(\boldsymbol{x}^i, y^i, f, \mathscr{D}, \Psi) = \phi\big(\Pr_{\mathcal{A}}[x^i_t = g_{\text{true}} \mid y^i, x^i_{-t}, \Psi]\big)$,

Using the reconstruction score $\mathcal{A}'$, we define the privacy leakage from reconstruction attacks.

Definition 4 (Imputation-Corrected Reconstruction Risk $\mathbb{L}^i_{\mathcal{A}}$)**.** For the training data $D_{\text{train}} \in \mathscr{D}$, let $\mathcal{A}'_{\text{in}}(\boldsymbol{x}^i, y^i) := \mathcal{A}'(\boldsymbol{x}^i, y^i, f, \mathscr{D}, \Psi \mid (\boldsymbol{x}^i, y^i) \in D_{\text{train}})$, and $\mathcal{A}'_{\text{out}}(\boldsymbol{x}^i, y^i) := \mathcal{A}'(\boldsymbol{x}^i, y^i, f, \mathscr{D}, \Psi \mid (\boldsymbol{x}^i, y^i) \notin D_{\text{train}})$, the per-person leakage on the attribute for person i is:

$$\mathbb{L}^i_{\mathcal{A}} := \Pr[\mathcal{A}'_{\text{in}}(\boldsymbol{x}^i, y^i) > \mathcal{A}'_{\text{out}}(\boldsymbol{x}^i, y^i)].$$

Importantly, Definition 4 adjusts for the improved imputation baseline and the per-person reconstruction difficulties simultaneously. First, $\mathcal{A}'_{\text{out}}(\boldsymbol{x}^i, y^i)$ captures the difficulty of imputing the individual's genotype when they are not seen in training. By comparing the $\mathcal{A}'_{\text{in}}(\boldsymbol{x}^i, y^i)$ to the $\mathcal{A}'_{\text{out}}(\boldsymbol{x}^i, y^i)$ of the same $\mathcal{A}$, the definition quantifies the privacy leakage caused by the inclusion of family members in training, removing the effect of having the family structure as the auxiliary information captured in imputation baseline. Second, $\mathcal{A}'_{\text{out}}(\boldsymbol{x}^i, y^i)$ captures the difficulty of imputing the genotype for person i; a person j will have a different imputation score. By adjusting $\mathcal{A}'_{\text{in}}(\boldsymbol{x}^i, y^i)$ with $\mathcal{A}'_{\text{out}}(\boldsymbol{x}^i, y^i)$ for person i, this definition accounts for the per-person reconstruction difficulty.

To empirically compute Definition 4, we adopt the same shadow model training paradigm as in (Carlini et al., 2022) to collect multiple $\mathcal{A}'_{\text{in}}(\boldsymbol{x}^i, y^i)$ and $\mathcal{A}'_{\text{out}}(\boldsymbol{x}^i, y^i)$ for each i, as in Algorithm 3 in Appendix A.6. We use Mann-Whitney U test to see if the reconstruction scores when person i is in the training is larger than the reconstruction scores when i is not. It is a non-parametric hypothesis test, hence, it does not assume the distributions of the reconstruction scores. Its effect size r_{MWU} measures how probable any element of confs_{in} is greater than $\text{confs}_{\text{out}}$, required by Definition 4.

When $r_{\text{MWU}} > 0.5$, it means $\Pr[\mathcal{A}'_{\text{in}}(\boldsymbol{x}^i, y^i) > \mathcal{A}'_{\text{out}}(\boldsymbol{x}^i, y^i)]$ is more than random chance; it is a positive signal where the reconstruction performance for training is greater than imputation, indicating there is a leakage. When $r_{\text{MWU}} \leq 0.5$, $\Pr[\mathcal{A}'_{\text{in}}(\boldsymbol{x}^i, y^i) > \mathcal{A}'_{\text{out}}(\boldsymbol{x}^i, y^i)]$ is below random chance; it means the reconstruction performance on training is no better than imputation, hence there is no leakage. It can happen when the model perfectly imputes the attribute (so no additional leakage occurs) or absolutely learns nothing about the attribute.

3.5 FAMILY-AWARE PRIVACY-PROTECTION MECHANISMS

We consider a typical data curation and model training setup of hospitals: (i) family information is collected, the family correlation is hence known a priori, and (ii) the collected families are usually

well-bounded, and the kinship between different families is negligible. That is, knowing one family does not give additional information about other families (i.e., the families are independent).

In genomics datasets, the correlated relatives of a target individual can cause elevated privacy leakage for that individual; the unit for privacy protection therefore should not be per person anymore. Instead, we should treat the entire family as an atomic unit and protect the privacy of the family as a whole.

The default framework to consider is group-DP, as described in Appendix A.4. Each group contains data points of a family, capped by max size $k_{\max}$. The model is trained with the standard DP-SGD algorithm with a record-level privacy budget ϵ_r (defined in Definition 1). The group privacy budget $\epsilon_g = k_{\max}\epsilon_r$ to account for the worst-case privacy leakage of the largest group. However, if there are groups with size $k \ll k_{\max}$, their privacy leakage is $k\epsilon_r \ll \epsilon_g$, wasting the privacy budget. To use the privacy budget more efficiently, we apply client-DP, *where each client contains all data points from one family*:

Definition 5 ((ϵ_Ψ, δ)-Client-DP)**.** For any datasets $X, X' \in \mathscr{D}$ differing by all records from one client (i.e., one family in our case), an algorithm M satisfies (ϵ_Ψ, δ)-client-DP if: $\Pr[M(X) \in S] \leq e^{\epsilon_\Psi} \Pr\left[M\left(X'\right) \in S\right] + \delta$. The client-DP ML training algorithm is summarized in Appendix A.5.

With client-DP, the privacy leakage of the entire family is bounded by ϵ_Ψ, and in turn, every family member's leakage is bounded by ϵ_Ψ.

4 RESULTS

In this section, we first outline the experimental setup, detailing the genomic datasets, family structure configurations, target models, and the prediction task. We first visually demonstrate how the proposed metric, imputation-corrected reconstruction risk quantifies the privacy leakage for both GENOFA and naive attacks. Then we show that under the new metric, GENOFA can elevate the privacy leakage of an individual by leveraging the Mendelian correlation. Hence, we define the privacy granularity as all data points from the family, and apply client-DP training algorithm; the results are compared with group-DP to show its superior privacy-utility tradeoff. Lastly, we empirically quantify the privacy leakage after training the models with client-DP, showing that the privacy leakage is significantly reduced.

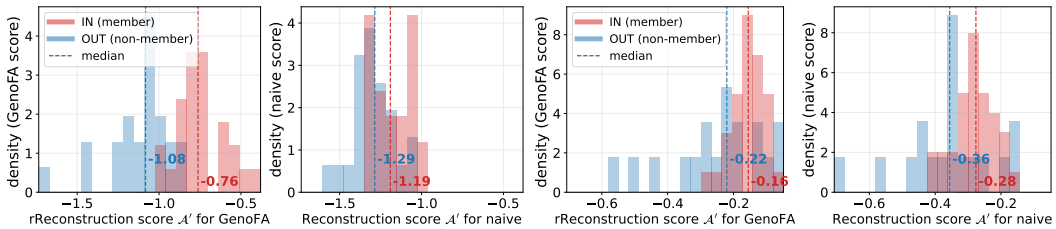
4.1 EXPERIMENTAL SETUP

Datasets and family structure. We extract genomic sequences from the 1000 Genomes Project (1000G) (Byrska-Bishop et al., 2022), anchoring our extraction on specific variant locations to capture diverse genomic distributions. Our attack focuses on first-degree relatives, reflecting standard screening practices in large-scale genomic studies. To evaluate the impact of kinship size on privacy leakage, we analyze two configurations: *small families* consisting exclusively of real 1000G trios ($\sim$3 individuals/family); *large families* are augmented with synthetic offspring ($\sim$12 individuals/family). We use the Mendelian transmission probabilities and a constant recombination rate with the Haldane mapping function to simulate the offspring genomes from real parents' genomes (from 1000G dataset). Full details on genomic range selection, preprocessing, and the simulation of synthetic offspring, and model training hyperparameters are provided in Appendix A.8.

Prediction task and model architectures. We evaluate two model architectures: a gLM, Nucleotide Transformer (NT) (Dalla-Torre et al., 2025) fine-tuned on our sequences, and a multi-layer perceptron (MLP) trained on tabular genomic variant encodings. Both models are trained to predict the mean Combined Annotation Dependent Depletion (CADD) score (Kircher et al., 2014) of the input sequence. CADD serves as a quantitative metric of variant deleteriousness. While CADD is a curated label rather than a measured clinical phenotype, it provides a reproducible evaluation target that induces the complex, non-linear genotype-to-label mapping required for disease risk prediction. This controlled setting isolates the privacy leakage attributable specifically to familial linkage. A detailed justification of the CADD objective and scaling is available in Appendix A.9.

4.2 FAMILY STRUCTURE ELEVATES PER-PERSON LEAKAGE $\mathbb{L}^i$

Recall that we use the reconstruction probability to define the reconstruction scores $\mathcal{A}'_{\text{in}}(\boldsymbol{x}^i, y^i)$'s and $\mathcal{A}'_{\text{out}}(\boldsymbol{x}^i, y^i)$'s (Definition 3). We further define $\mathbb{L}^i_{\mathcal{A}}$, the imputation-corrected reconstruction risk for person i using the effect size of the Mann-Whitney U test on person i's reconstruction scores (Definition 4). In this section, we first compare the reconstruction scores for two individuals. This



(a) $\mathbb{L}^i_{\text{GenoFA}} = 0.97 > \mathbb{L}^i_{\text{Naive}} = 0.75$. (b) $\mathbb{L}^i_{\text{GenoFA}} = \mathbb{L}^i_{\text{Naive}} = 0.72$.

Figure 1: **GENOFA can improve both the training and non-training reconstruction scores; privacy leakage increases only when the separation between $\mathcal{A}'_{\text{in}}(\boldsymbol{x}^i, y^i)$ and $\mathcal{A}'_{\text{out}}(\boldsymbol{x}^i, y^i)$ increases.** The median for each distribution is annotated in plots. In Figure 1a, GENOFA increases privacy leakage ($\mathbb{L}^i_{\text{GenoFA}} > \mathbb{L}^i_{\text{Naive}}$), visually reflected by the increasing separation between the two distributions. This is because GENOFA improves $\mathcal{A}'_{\text{in}}(\boldsymbol{x}^i, y^i)$'s twice as much as that of $\mathcal{A}'_{\text{out}}(\boldsymbol{x}^i, y^i)$. In Figure 1b, GENOFA increases $\mathcal{A}'_{\text{in}}(\boldsymbol{x}^i, y^i)$ and $\mathcal{A}'_{\text{out}}(\boldsymbol{x}^i, y^i)$ by similar amounts, hence, the privacy leakage $\mathbb{L}^i$ does not increase. Experiments use 50 shadow models (NT) and rs53576 (large) dataset. Similar effects are observed for MLP models as well, shown in Appendix A.13.

comparison visualizes when an individual may or may not experience an increased privacy leakage when family correlation is exploited.

Figure 1a shows that by leveraging the familial correlation, GENOFA increases the privacy leakage of this individual from $\mathbb{L}^i_{\text{Naive}} = 0.75$ to $\mathbb{L}^i_{\text{GenoFA}} = 0.97$. However, an increase in $\mathcal{A}'_{\text{in}}(\boldsymbol{x}^i, y^i)$ may not always mean an increase in privacy leakage; Figure 1b shows an unchanged privacy leakage ($\mathbb{L}^i_{\text{GenoFA}} = \mathbb{L}^i_{\text{Naive}} = 0.72$). Here we see that if the improved imputation baseline is ignored, one may conclude privacy leakage increases since the reconstruction score of training increases.

To see privacy leakage of GENOFA and naive for all individuals, we plot every individual's ($\mathbb{L}^i_{\text{GenoFA}}, \mathbb{L}^i_{\text{Naive}}$) pair as one dot in Figure 2a. Most individuals cluster around the diagonal line, indicating that their $\mathbb{L}^i_{\text{GenoFA}} \approx \mathbb{L}^i_{\text{Naive}}$. Every individual above the diagonal line has $\mathbb{L}^i_{\text{GenoFA}} > \mathbb{L}^i_{\text{Naive}}$. It means they experience higher privacy leakage due to the inclusion of family members in the training.

Table 3: Average per-person leakage differences between $\mathbb{L}^i_{\text{GenoFA}}$ and $\mathbb{L}^i_{\text{Naive}}(\Delta)$ using rs53576 small family sizes.

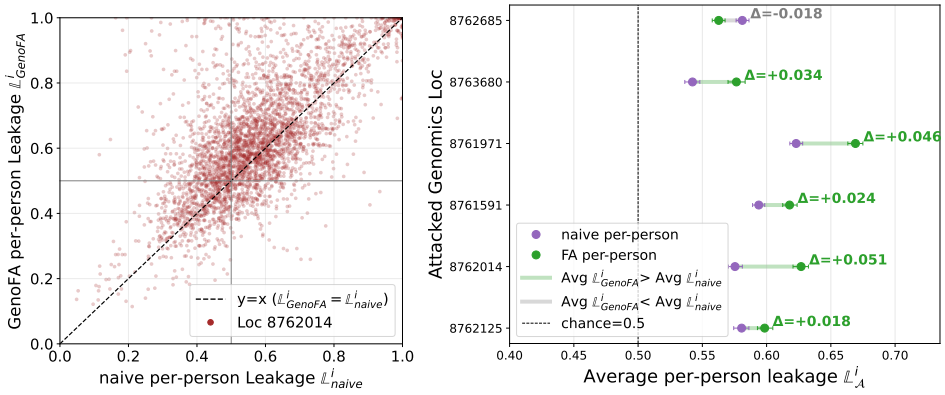
Genomics Loc (Chrom:Pos)	Δ
3:8762685	-0.003
3:8763680	+0.017
3:8761971	+0.008
3:8761591	+0.008
3:8762014	+0.023
3:8762125	+0.008

Figure 2b shows that GENOFA exposes higher privacy leakage on average for certain attacked attributes. For example, the privacy leakage of attribute 8762014 (the same attribute as in Figure 2a) is increased by $\Delta = 0.051$ on average. We notice that different attributes show different per-person privacy leakage. This is likely due to the same effect documented in (Yeom et al., 2018): if an attribute has strong or no influence on the label, the model will perfectly generalize or learn nothing on that attribute, respectively; hence there is no gap between the training and validation reconstruction scores, leading to no privacy leakage. If the attribute has small influence on the label, then the model cannot fully generalize but still learns some knowledge; here, the memorization effect dominates, leading to privacy leakage. Comparing with the attacks conducted on the datasets with smaller family sizes (Table 3), we also observe that GENOFA increases the privacy leakage more when family sizes are larger.

4.3 TRAINING WITH CLIENT-DP

In the previous section, we demonstrate GENOFA can increase per-person privacy leakage by leveraging familial correlation. Hence, the privacy granularity should not be at per-person level but at per-family level. As described in Section 3.5, group-DP naturally extends record-level DP by scaling the privacy granularity from per-record to per-group (in our case, per-family). Therefore, we use group-DP as a baseline and propose to use client-DP to improve the privacy-utility tradeoff.

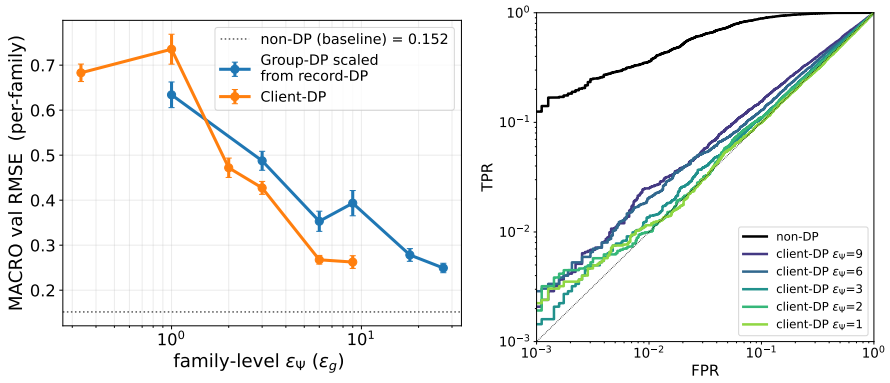
Figure 3a shows that client-DP in general gives higher utility at the same privacy budget than group-DP. Furthermore, it shows that a gLM fine-tuned under a modest privacy budget (e.g., $\epsilon_\Psi = 6.0$)



(a) $\mathbb{L}^i_{\text{GenoFA}}$ vs $\mathbb{L}^i_{\text{Naive}}$ (b) Average (across all persons) of $\mathbb{L}^i_{\text{GenoFA}}$ vs $\mathbb{L}^i_{\text{Naive}}$

Figure 2: **GENOFA leaks more privacy by leveraging the familial correlation.** Experiments use NT with rs53576 (large family). In Figure 2a, each data point is the pair ($\mathbb{L}^i_{\text{GenoFA}}$, $\mathbb{L}^i_{\text{Naive}}$) for one person. Figure 2b shows the average of $\mathbb{L}^i_{\text{GenoFA}}$ and $\mathbb{L}^i_{\text{Naive}}$ (across all individuals i). The confidence interval is calculated with bootstrapping of 1000 resamples. Δ quantifies the average increase on per-person leakage introduced by GENOFA. Results on different genomic locations, datasets, and models are shown in Appendices A.14, A.15, and A.16.

maintains strong model performance, achieving an RMSE of 0.27. This highlights the feasibility of training gLMs to retain practical utility while protecting the patients with rigorous DP guarantees. Results for different datasets and models are summarized in Appendix A.17.



(a) Client-DP has better Privacy-utility tradeoff (b) Per-family leakage is bounded after client-DP. AUC and TPR@0.1%FPR are in Table 5.

Figure 3: **Client-DP provides better privacy-utility tradeoff; the per-family privacy leakage is significantly reduced with client-DP.** Experiments use dataset rs53576 (small). In Figure 3a, the model utility is measured by per-family (macro) RMSE (lower is better). The error bar is the standard deviation across 5 runs using NT model. In Figure 3b we show the *leakage of the entire family* before and after DP using group-LiRA. Results on additional datasets are shown in Appendix A.18.

Quantifying the per-family privacy leakage. We now empirically evaluate the family-level privacy leakage. Note that in previous sections, individual-level privacy leakage has been estimated (i.e., $\mathbb{L}^i_{\text{GenoFA}}$). Here, the privacy is analyzed in a different granularity. In Figure 3b, we show empirical per-family privacy leakage using group-LiRA attack (Algorithm 4 in Appendix A.7). The non-DP models have the highest True Positive Rate (TPR) and the highest Area Under the Curve (AUC=0.968) over models trained with client-DP, showing that the privacy leakage is significantly reduced after applying the client-DP algorithm. At $\epsilon_\Psi = 6.0$, the attack success of models trained with client-DP

has an AUC of 0.531, which is more than 40% lower in comparison to non-DP models. Also, the AUC is close to 0.5 for all privacy budgets between $\epsilon_\Psi = 1.0$ and 9.0. This means the attack is close to random guessing, hence, little privacy is leaked empirically. It shows that for this specific prediction task, a single-digit privacy budget effectively constrains the empirical privacy risks.

Quantifying the per-person privacy leakage. In addition to reporting per-family privacy leakage (Figure 3b), we show the worst (max) per-person leakage after applying client-DP training mechanisms in Table 4 using the same setup. When no DP is applied, the leakage reaches its maximum value at 1.0. When the per-family privacy budget increases, the worst $\mathbb{L}^i_{\text{GenoFA}}$ privacy leakage also increases, but does not reach 1.0 as in the non-DP cases. For example, at $\epsilon_\Psi = 6$, the worst per-person leakage is 19.7% lower than the leakage for non-DP. It shows that when client-DP is applied, the per-family privacy leakage is bounded, and the leakage of the individuals in the family is also bounded. Results on different genomic locations and datasets are shown in A.19.

Table 4: **Worst per-person leakage is bounded after client-DP.** Attack location is 8761591 of rs53576 (small). 50 shadow models (MLP).

ϵ_Ψ	$\max_i \mathbb{L}^i_{\textbf{GenoFA}}$
1.0	0.7279
2.0	0.7639
3.0	0.7696
6.0	0.8030
9.0	0.8133
non-DP	1.0

5 Discussion

We demonstrate the usage of imputation-corrected reconstruction risk by quantifying privacy leakage under Mendelian correlation. This metric has the potential to be used in other domains where data points are correlated. The proposed client-DP training framework can also extend to scenarios where (i) all clients are independent, (ii) the correlation structure is known a-priori hence model trainers can group the data points into clients. This is helpful in cases where the correlation within each client often cannot be accurately modelled.

Limitations: Assuming family independence may yield overly large group sizes in bottlenecked populations. Additionally, our framework relies on publicly known variant locations, a standard genomic assumption, which breaks if these are private. While our inheritance modelling introduces simplifications suited for short genomic ranges (Appendix A.8), capturing complex biological relationships across longer sequences is left for future work. Lastly, identifying the specific individual and family profiles that maximize leakage or correlation-induced risks (a crucial step for privacy auditing) remains an open question for future attack evaluations.

Beyond Mendelian correlation: In addition to Mendelian correlation (a form of *value dependency* where data points are statistically linked via biological mechanisms), genomic studies frequently introduce *sampling dependency*. This arises when training cohorts are purposefully curated using biological groupings. Crucially, this sampling dependency is independent of value dependency, and the associated cohort curation metadata is typically released alongside the model, hence accessible to adversaries. In Section A.1, we apply group-LiRA to attack models trained with sampling dependency, demonstrating that exploiting these cohort structures leads to higher attack success.

6 Conclusion

In this work, we characterize privacy vulnerabilities arising from family-correlated data in gLMs and conventional MLP models. We introduce GenoFA, a family-aware genotype reconstruction attack that exploits Mendelian inheritance, showing that familial correlations amplify leakage whether relatives are in the training set or used as auxiliary information. To distinguish memorization from the inherent imputation capabilities of gLMs, we propose imputation-corrected reconstruction risk, which quantifies member–non-member separation while accounting for individual reconstruction difficulty. Our results demonstrate correlation-induced leakage and establish the family as the atomic unit of privacy protection. We further leverage a client-DP framework that bounds both family- and individual-level leakage while improving the privacy–utility trade-off over group-DP baselines. The framework reduces to person-level privacy for singleton families and extends to other dependency structures, including sampling dependencies in clinical cohort curation. As gLMs enable increasingly large-scale inference from correlated genomic data, privacy protections must explicitly account for these dependencies to remain robust in real-world applications.

AI USE STATEMENT

We utilized large language models (LLMs) to assist with coding and writing tasks. Specifically, LLMs were used to help extend the attack framework from attacking one attribute to attacking multiple attributes per dataset, adapt the attack pipeline from gLM to MLP architectures, optimize GPU memory usage, develop scripts for data extraction from genomic databases. Additionally, LLMs assisted in writing Hidden Markov Model (HMM) functions and Mendelian transmission modelling, writing scripts to merge results, generating visualization scripts. All generated code was manually reviewed and verified by the authors. Furthermore, LLMs were used to refine the grammar and clarity of the manuscript and used for macro definitions for writing consistency. All resulting text was thoroughly reviewed and edited by the authors.

ETHICS STATEMENT

This study exclusively utilizes publicly accessible, de-identified datasets (1000G, the human reference genome, and ClinVar). All computational experiments adhere to standard data-use protocols, ensuring that patient privacy is fully protected.

REPRODUCIBILITY STATEMENT

We will release the code for data preparation, model training, GENOFA attack, and group-DP scripts. Detailed experimental setup on the dataset (genomic range, offspring simulation and model training hyperparameters) is summarized in Appendix A.8. The group-DP and client-DP training algorithms are summarized in Appendices A.3 and A.5. The proposed attack GENOFA is provided in Section 3.3 and Appendix A.2. Group-LiRA algorithm is summarized in Appendix A.4

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, oct 2016. doi: 10.1145/2976749.2978318. URL https://doi.org/10.1145/2976749.2978318.

John M. Abowd, Matthew J. Schneider, and Lars Vilhuber. Differential privacy applications to bayesian and linear mixed model estimation. *Journal of Privacy and Confidentiality*, 5(1), 2013. doi: 10.29012/jpc.v5i1.627. URL https://doi.org/10.29012/jpc.v5i1.627.

Sergey Aganezov, Stephanie M Yan, Daniela C Soto, Melanie Kirsche, Samantha Zarate, Pavel Avdeyev, Dylan J Taylor, Kishwar Shafin, Alaina Shumate, Chunlin Xiao, et al. A complete reference genome improves analysis of human genetic variation. *Science*, 376(6588):eabl3533, 2022.

Reem Al-Saidi, Erman Ayday, and Ziad Kobti. Vespa: Vulnerability-enhanced selective privacy preservation adaptation against nucleotide inference for genomic embeddings in large language models. In *ICISSP (1)*, pp. 396–409, 2026.

Nour Almadhoun, Erman Ayday, and Özgür Ulusoy. Differential privacy under dependent tuples—the case of genomic privacy. *Bioinformatics*, 36(6):1696–1703, 2020.

Kerem Ayoz, Erman Ayday, and A. Ercüment Çiçek. Genome reconstruction attacks against genomic data-sharing beacons. *CoRR*, abs/2001.08852, 2020. URL https://arxiv.org/abs/2001.08852.

Asia Belfiore, Jonathan Passerat-Palmbach, and Dmitrii Usynin. Biologically-informed hybrid membership inference attacks on generative genomic models, 2025. URL https://arxiv.org/abs/2511.07503.

Marta Byrska-Bishop, Uday S. Evani, Xuefang Zhao, Anna O. Basile, Haley J. Abel, Allison A. Regier, André Corvelo, Wayne E. Clarke, Rajeeva Musunuri, Kshithija Nagulapalli, Susan Fairley, Alexi Runnels, Lara Winterkorn, Ernesto Lowy, Evan E. Eichler, Jan O. Korbel, Charles Lee, Tobias

Marschall, Scott E. Devine, William T. Harvey, Weichen Zhou, Ryan E. Mills, Tobias Rausch, Sushant Kumar, Can Alkan, Fereydoun Hormozdiari, Zechen Chong, Yu Chen, Xiaofei Yang, Jiadong Lin, Mark B. Gerstein, Ye Kai, Qihui Zhu, Feyza Yilmaz, Chunlin Xiao, Paul Flicek, Soren Germer, Harrison Brand, Ira M. Hall, Michael E. Talkowski, Giuseppe Narzisi, and Michael C. Zody. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell*, 185(18):3426–3440.e19, September 2022. ISSN 0092-8674. doi: 10.1016/j.cell.2022.08.004. URL https://doi.org/10.1016/j.cell.2022.08.004.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650. USENIX Association, August 2021. ISBN 978-1-939133-24-3. URL https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914, 2022. doi: 10.1109/SP46214.2022.9833649.

Micaela E Consens, Cameron Dufault, Michael Wainberg, Duncan Forster, Mehran Karimzadeh, Hani Goodarzi, Fabian J Theis, Alan Moses, and Bo Wang. Transformers and genome language models. *Nature Machine Intelligence*, 7(3):346–362, 2025.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P De Almeida, Hassan Sirelkhatim, et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22(2):287–297, 2025.

Neil M Davies, Gibran Hemani, Jenae M Neiderhiser, Hilary C Martin, Melinda C Mills, Peter M Visscher, Loïc Yengo, Alexander Strudwick Young, and Matthew C Keller. The importance of family-based sampling for biobanks. *Nature*, 634(8035):795–803, 2024.

Cynthia Dwork. A firm foundation for private data analysis. *Commun. ACM*, 54(1):86–95, jan 2011. ISSN 0001-0782. doi: 10.1145/1866739.1866758. URL https://doi.org/10.1145/1866739.1866758.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, aug 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL https://doi.org/10.1561/0400000042.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin (eds.), *Theory of Cryptography*, pp. 265–284, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

Congyu Fang, Adam Dziedzic, Lin Zhang, Laura Oliva, Amol Verma, Fahad Razak, Nicolas Papernot, and Bo Wang. Decentralised, collaborative, and privacy-preserving machine learning for multi-hospital data. *eBioMedicine*, 101:105006, March 2024. ISSN 2352-3964. doi: 10.1016/j.ebiom.2024.105006. URL http://dx.doi.org/10.1016/j.ebiom.2024.105006.

Veniamin Fishman, Yuri Kuratov, Aleksei Shmelev, Maxim Petrov, Dmitry Penzar, Denis Shepelin, Nikolay Chekanov, Olga Kardymon, and Mikhail Burtsev. Gena-lm: a family of open-source foundational dna language models for long sequences. *Nucleic Acids Research*, 53(2):gkae1310, 2025.

Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An End-to-End case study of personalized warfarin dosing. In *23rd USENIX Security Symposium (USENIX Security 14)*, pp. 17–32, San Diego, CA, August 2014. USENIX Association. ISBN 978-1-931971-15-7. URL https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/fredrikson_matthew.

Sayan Ghosal, Youssef Barhomi, Tejaswini Ganapathi, Amy Krystosik, Lakshmi Krishnan, Sashidhar Guntury, Donghui Li, Alzheimer's Disease Neuroimaging Initiative, Francesco Paolo Casale, and Theofanis Karaletsos. Variantformer: A hierarchical transformer integrating dna sequences with genetic variations and regulatory landscapes for personalized gene expression prediction. *bioRxiv*, 2025.

A. Harmanci and M. Gerstein. Quantification of private information leakage from phenotype-genotype data: linking attacks. *Nature Methods*, 13(3):251–256, Mar 2016. doi: 10.1038/nmeth.3746.

Nils Homer, Szabolcs Szelinger, Matthew Redman, David Duggan, Waibhav Tembe, Waibhav Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David W Craig. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS Genetics*, 4(8):e1000167, 2008. doi: 10.1371/journal.pgen.1000167.

Laurence J Howe, Michel G Nivard, Tim T Morris, Ailin F Hansen, Humaira Rasheed, Yoonsu Cho, Geetha Chittoor, Rafael Ahlskog, Penelope A Lind, Teemu Palviainen, et al. Within-sibship genome-wide association analyses decrease bias in estimates of direct genetic effects. *Nature genetics*, 54(5):581–592, 2022.

Daniel Kifer and Ashwin Machanavajjhala. Pufferfish: A framework for mathematical privacy definitions. *ACM Trans. Database Syst.*, 39(1), January 2014. ISSN 0362-5915. doi: 10.1145/2514689. URL https://doi.org/10.1145/2514689.

Martin Kircher, Daniela M Witten, Preti Jain, Brian J O'roak, Gregory M Cooper, and Jay Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics*, 46(3):310–315, 2014.

Ao Lan, Yudi Pawitan, and Xia Shen. Genomic privacy risks in gwas summary statistics. November 2025.

Melissa J Landrum, Jennifer M Lee, Mark Benson, Garth R Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee Jang, et al. Clinvar: improving access to variant interpretations and supporting evidence. *Nucleic acids research*, 46(D1):D1062–D1067, 2018.

Martin Otto Lange. *Utility of Semantic Privacy Notions for Correlated Data*. Projecte final de màster oficial, UPC, Escola Tècnica Superior d'Enginyeria de Telecomunicació de Barcelona, Departament d'Enginyeria Telemàtica, 2024. URL https://hdl.handle.net/2117/422584.

Zehui Li, Vallijah Subasri, Guy-Bart Stan, Yiren Zhao, and Bo Wang. Gv-rep: A large-scale dataset for genetic variant representation learning. *Advances in Neural Information Processing Systems*, 37:134701–134720, 2024.

Changchang Liu, Supriyo Chakraborty, and Prateek Mittal. Dependence makes you vulnberable: Differential privacy under dependent tuples. In *NDSS*, volume 16, pp. 21–24, 2016.

Nicholas A Marston, Frederick K Kamanu, Giorgio EM Melloni, Gavin Schnitzler, Aaron Hakim, Rosa X Ma, Helen Kang, Daniel I Chasman, Robert P Giugliano, Patrick T Ellinor, et al. Endothelial cell-related genetic variants identify ldl cholesterol-sensitive individuals who derive greater benefit from aggressive lipid lowering. *Nature medicine*, 31(3):963–969, 2025.

Andres Munoz medina, Robert Istvan Busa-Fekete, Umar Syed, and Sergei Vassilvitskii. On the pitfalls of label differential privacy. In *NeurIPS 2021 Workshop LatinX in AI*, 2021. URL https://openreview.net/forum?id=2sWidqliCDG.

Alexander Nemecek, Wenbiao Li, Xiaoqian Jiang, Jaideep Vaidya, and Erman Ayday. *Quantifying Memorization and Privacy Risks in Genomic Language Models*. Association for Computing Machinery, New York, NY, USA, 2026. ISBN 9798400726538. URL https://doi.org/10.1145/3807503.3819470.

Bristena Oprisanu, Georgi Ganev, and Emiliano De Cristofaro. On utility and privacy in synthetic genomic data. *arXiv preprint arXiv:2102.03314*, 2021.

Brooks Paige, James Bell, Aurélien Bellet, Adrià Gascón, and Daphne Ezer. Reconstructing genotypes in private genomic databases from genetic risk scores. In Russell Schwartz (ed.), *Research in Computational Molecular Biology*, pp. 266–268, Cham, 2020. Springer International Publishing.

Valentin Roth and Marco Avella-Medina. Differential privacy with dependent data, 2025. URL https://arxiv.org/abs/2511.18583.

Lynn Y Sakai, Douglas R Keene, Marjolijn Renard, and Julie De Backer. Fbn1: The disease-causing gene for marfan syndrome and other genetic disorders. *Gene*, 591(1):279–291, 2016.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership Inference Attacks Against Machine Learning Models . In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18, Los Alamitos, CA, USA, May 2017. IEEE Computer Society. doi: 10.1109/SP.2017.41. URL https://doi.ieeecomputersociety.org/10.1109/SP.2017.41.

Suyash S. Shringarpure and Carlos D. Bustamante. Privacy risks from genomic data-sharing beacons. *The American Journal of Human Genetics*, 97(5):631–646, 2026/05/22 2015.

Thomas Steinke, Milad Nasr, and Matthew Jagielski. Privacy auditing with one (1) training run. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 49268–49280. Curran Associates, Inc., 2023.

M. Thomas, N. Mackes, A. Preuss-Dodhy, T. Wieland, and M. Bundschus. Assessing privacy vulnerabilities in genetic data sets: Scoping review. *JMIR Bioinformatics and Biotechnology*, 5:e54332, 2024. doi: 10.2196/54332. URL https://bioinform.jmir.org/2024/1/e54332.

Caroline Uhler, Aleksandra B. Slavković, and Stephen E. Fienberg. Privacy-preserving data sharing for genome-wide association studies. *Journal of Privacy and Confidentiality*, 5(1), 2013. doi: 10.29012/jpc.v5i1.629. URL https://doi.org/10.29012/jpc.v5i1.629.

Nora von Thenen, Erman Ayday, and A Ercument Cicek. Re-identification of individuals in genomic data-sharing beacons via allele inference. *Bioinformatics*, 35(3):365–371, 02 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty643. URL https://doi.org/10.1093/bioinformatics/bty643.

Bin Yang, Issei Sato, and Hiroshi Nakagawa. Bayesian differential privacy on correlated data. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, SIGMOD '15, pp. 747–762, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450327589. doi: 10.1145/2723372.2747643. URL https://doi.org/10.1145/2723372.2747643.

Yiqi Yao and Alejandro Ochoa. Limitations of principal components in quantitative genetic association models for human studies. *eLife*, 12:e79238, may 2023. ISSN 2050-084X. doi: 10.7554/eLife.79238. URL https://doi.org/10.7554/eLife.79238.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting . In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pp. 268–282, Los Alamitos, CA, USA, July 2018. IEEE Computer Society. doi: 10.1109/CSF.2018.00027. URL https://doi.ieeecomputersociety.org/10.1109/CSF.2018.00027.

Emre Yilmaz, Tianxi Ji, Erman Ayday, and Pan Li. Genomic data sharing under dependent local differential privacy, 2021. URL https://arxiv.org/abs/2102.07357.

Johannes Zschocke, Peter H Byers, and Andrew OM Wilkie. Mendelian inheritance revisited: dominance and recessiveness in medical genetics. *Nature Reviews Genetics*, 24(7):442–463, 2023.

A Appendix

A.1 Sampling Dependency Amplifies Privacy Leakage

The preceding sections examined family correlation, a form of *value dependency* where records are statistically linked through underlying biological mechanisms (e.g., Mendelian inheritance). However, genomic studies also introduce a distinct vulnerability arising from cohort curation: *sampling dependency*.

Sampling dependency occurs when records are included or excluded together based on biological groupings, independent of attribute correlations. For example, variant-effect studies routinely hold out entire genes to test model generalization (Landrum et al., 2018), while family-based studies jointly enroll all relatives (Howe et al., 2022; Davies et al., 2024). Consequently, two unrelated patients whose variants map to the same held-out gene share perfect sampling dependency despite having zero value dependency; the inclusion of one record deterministically implies the inclusion of the other. This practice is standard in studies using datasets like ClinVar to ensure proper model evaluation (see Appendix A.10).

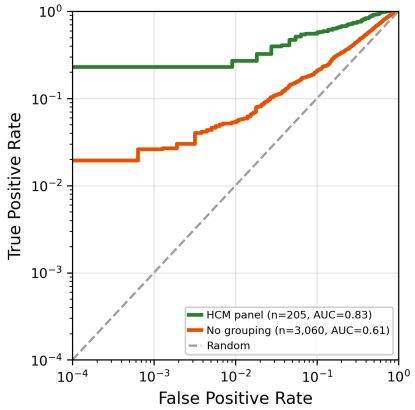


Figure 4: **Exploiting sampling dependency improves membership inference attacks.** Group-LiRA, which exploits the HCM gene panel structure (green, 205 gene groups), substantially outperforms standard variant-level LiRA without grouping (orange, 3,060 variants). Because curation protocols are routinely published for reproducibility, adversaries can readily exploit this exposed structure.

Under a realistic threat model where an adversary leverages these published curation protocols to execute group-level membership inference, the privacy risks escalate significantly. Applying Group-LiRA to ClinVar reveals that grouping records by HCM gene panel substantially amplifies attack success compared to independent sampling (Figure 4). This demonstrates that gene-level curation creates exploitable sampling dependencies, even among strictly unrelated individuals. Crucially, while Group-LiRA exploits *sampling* dependency and the family-aware reconstruction attack (Section 3) exploits *value* dependency, both phenomena independently violate the independent-record assumption of standard differential privacy. Mitigating either vulnerability requires treating the dependent group as the fundamental unit of protection via group- or client-level DP.

A.2 Mendelian Transmission and Recombination $T(k \mid m, d)$

'Recall we want to find

$$T(k \mid m, d) := \Pr[x_t^{\text{kid}} = k \mid x_t^{\text{mom}} = m, x_t^{\text{dad}} = d, \mathbf{x}_{-t}], \sum_k T(k \mid m, d) = 1.$$

For simplicity, we ignored all random mutations (de novo mutations) happened to the kids. For simplicity, we ignored the hapmap and assumed the recombination or crossover happens ρ_l based on Haldane mapping function. ρ_l depends on how far apart two genomic locations are.

Everyone has two haplotypes. Let $\{A, B\}$ be the indices of the two haplotypes.

Let *crossover indicator* $r = (r_1, \ldots, r_L) \in \{A, B\}^L$. Each entry $r_\ell \in \{A, B\}$ indicates which of parent's two strands is transmitted to the child at locus ℓ. For example, $r = (A, A, B)$ means this parent transmitted the genotype from H_A for the first two positions and switched to the genotype

from H_B for the third position to the kid. The rate at where the switching is the recombination rate ρ_l. The probability of a crossover indicator is:

$$\Pr[r] = \tfrac{1}{2} \prod_{\ell=1}^{L-1} \big[(1-\rho_\ell)\,\mathbb{1}[r_{\ell+1} = r_\ell] + \rho_\ell\,\mathbb{1}[r_{\ell+1} \neq r_\ell]\big]. \tag{4}$$

Let each parent p has two phased strands $\big(x^p_{A,\cdot}, x^p_{B,\cdot}\big)$, where $x^p_{s,\ell}$ denotes the allele on strand $s \in \{A, B\}$ at locus ℓ. let x^k_s be any strand of the (candidate) kid's genome.

$$F^p(x^k_s) \;:=\; \sum_{r \in \{A,B\}^L} \Pr[r] \prod_{\ell=1}^{L} \mathbb{1}\Big[x_\ell = x^p_{r_\ell,\ell}\Big] \;=\; \Pr[\text{parent } p \text{ transmits } x^k_s], \tag{5}$$

In 1000G data, the parent of origin for each haplotype is not provided. We need to infer it by matching the genomes within the family if family structure is provided. If kid's haplotype A (x^k_A) is inherited from mother and kid's haplotype B (x^k_B) is inherited from dad, the two haplotypes are conditionally independent and hence:

$$\mathbb{P}_{A \leftarrow \texttt{mom}, B \leftarrow \texttt{dad}}\big[x^{\text{kid}}_A, x^{\text{kid}}_B \mid x^{\text{mom}}, x^{\text{dad}}\big] = F^{\text{mom}}(x^{\text{kid}}_A) F^{\text{dad}}(x^{\text{kid}}_B), \tag{6}$$

Let $\phi \in \{0, 1\}$ represent the prior: $\phi = 0$ when $A \leftarrow$ mom, $B \leftarrow$ dad; $\phi = 1$ when $B \leftarrow$ mom, $A \leftarrow$ dad. $\Pr[\phi = 0] = \Pr[\phi = 1] = 0.5$

$$\mathbb{P}\big[x^{\text{kid}}_A, x^{\text{kid}}_B \mid x^{\text{mom}}, x^{\text{dad}}\big] = 0.5 \cdot F^{\text{mom}}(x^{\text{kid}}_A) F^{\text{dad}}(x^{\text{kid}}_B) + 0.5 \cdot F^{\text{mom}}(x^{\text{kid}}_B) F^{\text{dad}}(x^{\text{kid}}_A), \tag{7}$$

$$\begin{aligned}
\mathbb{P}\big[x^{\text{kid}}_A, x^{\text{kid}}_B \mid x^{\text{mom}}, x^{\text{dad}}, x^{\text{kid}}_{-t}\big] &= \frac{\mathbb{P}\big[x^{\text{kid}}_{A,t}, x^{\text{kid}}_{B,t}, x^{\text{kid}}_{-t} \mid x^{\text{mom}}, x^{\text{dad}}\big]}{\sum_{x^{\text{kid}}_{A,t}\prime, x^{\text{kid}}_{B,t}\prime} \mathbb{P}\big[x^{\text{kid}}_{A,t}\prime, x^{\text{kid}}_{B,t}\prime, x^{\text{kid}}_{-t} \mid x^{\text{mom}}, x^{\text{dad}}\big]} \\
\mathbb{P}\big[x^{\text{kid}}_{A,t}, x^{\text{kid}}_{B,t} \mid x^{\text{mom}}, x^{\text{dad}}, x^{\text{kid}}_{-t}\big] &= \frac{0.5 \cdot F^{\text{mom}}(x^{\text{kid}}_A) F^{\text{dad}}(x^{\text{kid}}_B) + 0.5 \cdot F^{\text{mom}}(x^{\text{kid}}_B) F^{\text{dad}}(x^{\text{kid}}_A)}{\sum_{x^{\text{kid}}_{A,t}\prime, x^{\text{kid}}_{B,t}\prime} 0.5 \cdot F^{\text{mom}}(x^{\text{kid}}_A) F^{\text{dad}}(x^{\text{kid}}_B) + 0.5 \cdot F^{\text{mom}}(x^{\text{kid}}_B) F^{\text{dad}}(x^{\text{kid}}_A)} \\
&= \frac{F^{\text{mom}}(x^{\text{kid}}_A) F^{\text{dad}}(x^{\text{kid}}_B) + F^{\text{mom}}(x^{\text{kid}}_B) F^{\text{dad}}(x^{\text{kid}}_A)}{\sum_{x^{\text{kid}}_{A,t}\prime, x^{\text{kid}}_{B,t}\prime} F^{\text{mom}}(x^{\text{kid}}_A) F^{\text{dad}}(x^{\text{kid}}_B) + F^{\text{mom}}(x^{\text{kid}}_B) F^{\text{dad}}(x^{\text{kid}}_A)}
\end{aligned} \tag{8}$$

$$\boxed{T(k \mid m, d) :=\; \Pr[x^{\text{kid}}_t = k \mid x^{\text{mom}}_t = m, x^{\text{dad}}_t = d, \mathbf{x}_{-t}] = \mathbb{P}\big[x^{\text{kid}}_{A,t}, x^{\text{kid}}_{B,t} \mid x^{\text{mom}}, x^{\text{dad}}, x^{\text{kid}}_{-t}\big]} \tag{9}$$

Family-aware reconstruction likelihood if multiple kids per family If the family contains the parents and multiple kids, then the kids are conditionally independent given the parents:

$$\Pr[k_1, \ldots, k_K, m, d \mid \mathbf{x}_{-t}] \;=\; \pi_2(m)\,\pi_3(d) \prod_{i=1}^{K} T_i(k_i \mid m, d) \tag{10}$$

Let $\boldsymbol{y} = (y^{k_1}, ..., y^{k_K}, y^m, y^d)$ be the labels for kid(s), mother and father of the family.

$$\begin{aligned}
\Pr[k_1, ..., k_K, m, d \mid \boldsymbol{y}, \mathbf{x}_{-t}] &\propto \Pr[\boldsymbol{y} \mid k_1, ..., k_K, m, d, \mathbf{x}_{-t}]\;\Pr[k_1, ..., k_K, m, d \mid \mathbf{x}_{-t}] \\
&= \Big[\prod_{i=1}^{K} L_i(k_i) L_2(m) L_3(d)\Big]\;\Pr[k_1, ..., k_K, m, d \mid \mathbf{x}_{-t}] \\
&= \Big[\prod_{i=1}^{K} L_i(k_i) L_2(m) L_3(d)\Big]\;\pi_2(m)\,\pi_3(d) \prod_{i=1}^{K} T_i(k_i \mid m, d)
\end{aligned}$$

$$\text{Let} \quad W(k_1, \ldots, k_K, m, d) = \;\pi_2(m)\pi_3(d)\,L_2(m)L_3(d) \prod_{i=1}^{K} T_i(k_i \mid m, d)\,L_i(k_i),$$

Then we normalize W and marginalize over all people's genotypes other than the target person to find the probability of GENOFA attack for the target person.

A.3 DP-SGD

DP-SGD algorithm is described in Algorithm 1.

Algorithm 1 DP-SGD (Abadi et al., 2016)

Require: Dataset X, Mini-Batch Size B, Clipping Norm C, Noise Multiplier σ, Model Parameters W, Loss Function L, Learning Rate η

$W_0 \leftarrow$ `RandomInit()`
for $t \leftarrow 0, \ldots, T-1$ **do**
  `Sample Mini-batch from` X `with Mini-batch size` B
  for $x_b \in$ `Mini-batch` **do**
    $g_b = \nabla_W L(W_t, x_b)$ ▷ *Per-example gradient calculation*
    $\bar{g_b} = g_b / \max(C^{-1}||g_b||_2, 1)$ ▷ *Per-example gradient clipping*
  end for
  $g = \frac{1}{||B||}(\sum_b \bar{g_b} + \mathcal{N}(0, (C\sigma)^2))$ ▷ *Add calibrated Gaussian Noise*
  $W_{t+1} \leftarrow W_t - \eta g$
end for

return W_T

A.4 GROUP-DP

Group-DP is described in Definition 6.

Definition 6. Group DP: For any datasets $X, X' \in \mathscr{X}$ differing by k units ($||X - X'|| \leq k$), an $(\epsilon, \delta)-$DP algorithm is $(k\epsilon, ke^{(k-1)\epsilon}\delta)$-DP: $\Pr[M(X) \in S] \leq e^{k\epsilon} \Pr[M(X') \in S] + ke^{(k-1)\epsilon}\delta$.

Group-DP protects against the worst-case dependency structure, i.e., data points within the group has perfect correlation. When applying group-DP, the models can be trained with record-level DP (DP-SGD (Abadi et al., 2016) (Algorithm 1)) and scale the group privacy budget based on the number of data records contributed by the same group post hoc.

A.5 CLIENT-DP

Client-DP algorithm is described in Algorithm 2.

Algorithm 2 Client DP

Require: Number of clients N, model state W_t, Clipping Norm C, Noise Multiplier σ, subsampling rate p, target privacy budget ϵ, total rounds T

$W_0 \leftarrow$ `RandomInitialization()`
for $t \leftarrow 0, \ldots, T-1$ **do**
  Randomly select a subset of clients L_t with probability p (Poisson subsampling)
  for $i \in L_t$ **do**
    $\Delta W_t^i = \texttt{Model_Update}(W_{t-1}, X_i)$
    $\Delta\overline{W_t^i} = \Delta W_t^i / \max\left(\| C^{-1}\Delta W_t^i\|_2, 1\right)$ ▷ *Clip per-client model update*
  end for
  $\Delta\widetilde{W}_t = \frac{1}{pN}\left(\sum_{i\in L_t} \Delta\overline{W_t^i} + \mathcal{N}(0, (C\sigma)^2)\right)$ ▷ *Aggregate and add noise*
  $W_t = W_{t-1} + \text{Update_Rule}(\Delta\widetilde{W}_t)$
end for

return W_T

A.6 SHADOW MODEL TRAINING FOR IMPUTATION-CORRECTED RECONSTRUCTION RISK

The detailed shadow model training algorithm is presented in Algorithm 3

Algorithm 3 Quantification of attribute inference advantage for the target person. The shadow models are trained with or without the target person (and their family members).

Require: model f,target person $(\boldsymbol{x}^i, y^i)$, the family members of target person $\{(\boldsymbol{x}^j, y^j)\}$, data distribution $\mathscr{D}$, attack method $\mathcal{A} \in \{\texttt{naive}, \textsc{GenoFA}\}$

$\text{confs}_{\text{in}} = \{\}$
$\text{confs}_{\text{out}} = \{\}$
for N times **do**
  $D_{\text{attack}} \xleftarrow{\$} \mathscr{D}$ ▷ *Sample a shadow dataset*
  $f_{\text{in}} \leftarrow \mathcal{T}(D_{\text{attack}} \cup \{(\boldsymbol{x}^i, y^i), (\boldsymbol{x}^j, y^j)\})$ ▷ *train IN model*
  $\text{confs}_{\text{in}} \leftarrow \text{confs}_{\text{in}} \cup \mathcal{A}'(\boldsymbol{x}^i, y^i, f_{\text{in}}, \mathscr{D}, \Psi)$ ▷ *Reconstruction scores for IN as in Definition 3*
  $f_{\text{out}} \leftarrow \mathcal{T}(D_{\text{attack}} \backslash \{(\boldsymbol{x}^i, y^i), (\boldsymbol{x}^j, y^j)\})$ ▷ *train OUT model*
  $\text{confs}_{\text{out}} \leftarrow \text{confs}_{\text{out}} \cup \mathcal{A}'(\boldsymbol{x}^i, y^i, f_{\text{out}}, \mathscr{D}, \Psi)$ ▷ *Reconstruction scores for OUT as in Definition 3*
end for
return $\mathbb{L}^i_{\mathcal{A}} = \text{Leakage}(\boldsymbol{x}^i, y^i \mid \mathcal{A}) = r_{\text{MWU}}(\text{confs}_{\text{in}}, \text{confs}_{\text{out}})$ ▷ *effect size of Mann-Whitney U test*

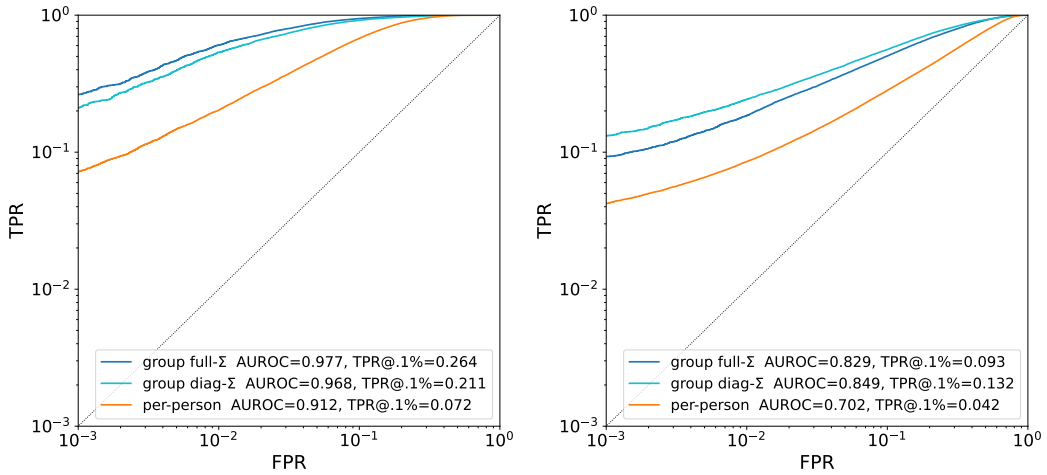
A.7 GROUP-LIRA

Likelihood Ratio Attack (LiRA) (Carlini et al., 2022) is a membership inference attack where the adversary is trying to find out whether a specific data point is used in training dataset or not. Adapting the LiRA algorithm, we propose Group-LiRA, as described in Algorithm 4. Here, the adversary is trying to find if the entire group of data points are in training or not. Following the convention (Carlini et al., 2022), we plot the TPR vs FPR to visualize the success rate. As a sanity check, we plot TPR vs FPR for both the group-LiRA together with regular record-level LiRA. in Figure 5. We can see that group-LiRA have consistently higher TPR than record-LiRA at all FPR rates, showing the group-LiRA is more successful than regular record level LiRA; it also means that it is easier for the adversary to find out the membership information about one group of data points than one single data points, meaning a group of data points is leaking more privacy than one single data point. Also, it is worth noting that when calculating the covariance matrices Σ_{in} Σ_{out}, when the group size is large, using a diagonal matrix to approximate it as oppose to using the full matrix can generate more successful attacks. When group size is small, then using the full matrix leads to higher attack success. In later experiments, we will use diagonal matrix approximations in group-LiRA. For regression tasks, we use the log mean squared error of each data point as the scores, i.e., $\log(\text{MSE}(f(x_k), y_k)) = \phi(f(x_k)_{y_k})$.

Algorithm 4 Group Likelihood Ratio Attack (Group-LiRA) (adapted from (Carlini et al., 2022)). The shadow models are trained with or without the target group of examples to estimate mean and covariance of the loss distributions, and compute a likelihood ratio test for all data points in the group.)

Require: model f, group of K examples $\{(x_k, y_k)\}$, data distribution $\mathscr{D}$

$\text{confs}_{\text{in}} = \{\}$
$\text{confs}_{\text{out}} = \{\}$
for N times **do**
  $D_{\text{attack}} \xleftarrow{\$} \mathscr{D}$ ▷ *Sample a shadow dataset*
  $f_{\text{in}} \leftarrow \mathcal{T}(D_{\text{attack}} \cup \{(x_k, y_k)\})$ ▷ *train IN model*
  $\text{confs}_{\text{in}} \leftarrow \text{confs}_{\text{in}} \cup \{\phi(f_{\text{in}}(x_k)_{y_k})\}$
  $f_{\text{out}} \leftarrow \mathcal{T}(D_{\text{attack}} \backslash \{(x_k, y_k)\})$ ▷ *train OUT model*
  $\text{confs}_{\text{out}} \leftarrow \text{confs}_{\text{out}} \cup \{\phi(f_{\text{out}}(x_k)_{y_k})\}$
end for
$\mu_{\text{in}} \leftarrow \texttt{mean}(\text{confs}_{\text{in}})$
$\mu_{\text{out}} \leftarrow \texttt{mean}(\text{confs}_{\text{out}})$
$\Sigma_{\text{in}} \leftarrow \texttt{Cov}(\text{confs}_{\text{in}})$
$\Sigma_{\text{out}} \leftarrow \texttt{Cov}(\text{confs}_{\text{out}})$
$\text{conf}_{\text{obs}} = \phi(f(x_k)_{y_k})$ ▷ *query target model*
return $\Lambda = \dfrac{p(\text{conf}_{\text{obs}} \mid \mathcal{N}(\mu_{\text{in}}, \Sigma_{\text{in}}))}{p(\text{conf}_{\text{obs}} \mid \mathcal{N}(\mu_{\text{out}}, \Sigma_{\text{out}}))}$ ▷ *same score for the K examples in the group*



(a) rs53576 (small) group size ≤ 6

(b) rs53576 (large) group size ≤ 14

Figure 5: **Group-LiRA is more successful than regular per-record LiRA, meaning group of data points is leaking more information than a single data point.** Plotted with 256 shadow models with MLP. Group-LiRA have higher TPR than record-LiRA at all FPR rates; the adversary can distinguish the membership information for the entire group much better than a single data point. It means that the group of data points is leaking more privacy than one single data point, which is consistent with the intuition. Also, we plot the attack performance of estimating the the covariance matrices Σ_{in} Σ_{out} using both the full matrix and diagonal matrix. We observe that when the group size is large, using a diagonal matrix lead to more successful attacks; when group size is small, full matrix lead to more successful attack. We will use diagonam matrix in later experiments unless otherwise stated.

A.8 Detailed Experimental Setup

Genomic range selection and preprocessing. For each dataset, we anchored on a focal variant location and extracted a 2,801 base-pair (bp) window. We selected four anchor variant locations across two chromosomes (rs53576 and rs237890 on Chr 3; rs2291844 and rs10838549 on Chr 11) to capture diverse linkage disequilibrium structures and minor allele frequencies. We present results for the dataset centred at rs53576 variant location in the main text; additional datasets centred at different locations are in Appendix A.15, A.16, A.17, A.18, and A.19. During preprocessing, both phased haplotypes per individual were concatenated and separated by an `<unk>` token to fill the Nucleotide Transformer's 6,000-token context window.

Simulated offspring generation. The 1000G dataset typically contains only one child per trio. To create the large families dataset (around 12 individuals/family), we generated synthetic offspring by sampling real parental haplotypes using Mendelian transmission probabilities and a constant recombination rate of 10,000 cM/Mb with the Haldane mapping function. This rate is substantially higher than the typical human average (around 1 cM/Mb) to ensure sufficient recombination diversity within the relatively small genomic windows analyzed ($\tilde{2}$,800 bp). While HapMap-derived rates would be more biologically accurate, they would produce negligible recombination at this scale, limiting offspring diversity.

Simplification when modelling genomics We make a few biological and computational choices in this work. First we omit *de novo* mutations and genotyping errors evaluates only error-free Mendelian transmissions. In the genomic ranges we sampled, the percentage of such cases is negligible. But for longer sequences or genomic range high in random mutation, the rate of random mutation shall also be modelled. Second, we only sample single nucleotide variants (SNVs) as the genomic variants in our dataset, omitting insertions or deletions. This is a common practice in genomics studies since SNVs are the most common type of variants. Finally, we omit explicit linkage disequilibrium (LD) conditioning due to the limited context lengths of these genomic windows of the gLM we

use. For longer sequences where the LD effect is more pronounced and documented, one can adopt frameworks like the Li-Stephens model to capture long-range LD across larger windows. Furthermore, because family structures in the 1000G primarily comprise trios, we focus on modelling Mendelian correlations for these parent-child trios rather than multi-generational extended pedigrees.

Model training hyperparameters. For Nucleotide Transformer v2 (NT), we use the 500M architecture pretrained on the human reference genome. For small-family datasets, we cap the max family size to be 3. Client-DP is trained with $(\epsilon_\Psi, \delta_\Psi = 10^{-5})$ for $(\epsilon_\Psi = [1/3, 1, 2, 3, 6, 9]$, group-DP is trained with record-DP $(\epsilon_r, \delta_r = 10^{-5})$ for $\epsilon_r = [1/3, 1, 2, 3, 6, 9]$, with group-DP $(\epsilon_g = k \cdot \epsilon_r, \delta_g = ke^{(k-1)\epsilon_r}\delta)$, with $k = 3$ (i.e., the max family size is capped at 3), hence $\epsilon_g = [1, 3, 6, 9, 18, 27]$. For MLP architecture, we do not cap the max family size. Client-DP is trained with $(\epsilon_\Psi, \delta_\Psi = 10^{-5})$ for $(\epsilon_\Psi = [1, 2, 3, 6, 9]$, group-DP is trained with record-DP $(\epsilon_r, \delta_r = 10^{-5})$ for $\epsilon_r = [1, 2, 3, 6, 9]$. The hyperparameters (e.g., learning rate, clipping norm for DP, number of epochs, sampling rate) are swept with one seed first, and then the best configuration is used to train with multiple seeds to collect shadow models and compute confidence intervals.

A.9 Pathogenicity Prediction Target Details

Both models predict the mean Combined Annotation Dependent Depletion (CADD) score, scaled using the Phred framework. CADD integrates over 60 genomic annotations into a single metric of variant deleteriousness, effectively acting as a phenotypic label for the genotype of a single variant location. We use the mean CADD score across all variants in the window to ensure every variant genotype contributes to the sequence's overall phenotypic label, providing a learnable, sequence-level objective for the model.Predicting the mean window score provides a robust evaluation target for three reasons: (1) it is computable directly from sequences without external databases, enabling reproducible benchmarking; (2) it induces the necessary non-linear genotype-to-label mapping required for disease risk prediction; and (3) CADD strongly correlates with clinical pathogenicity. Therefore, models that leak CADD-associated variants will inherently leak sensitive information when fine-tuned on clinical outcomes, establishing a strict lower bound for privacy risks in real-world deployments.

A.10 ClinVar Sampling Dependency Attacks

Threat Model Justification. Extending the standard LiRA framework to incorporate knowledge of the sampling strategy is highly practical in medical research, where cohort curation protocols are typically published to ensure reproducibility. This transparency enables an adversary with black-box model access to infer whether a higher-order unit (e.g., all variants associated with a specific gene) was included in the training cohort, rather than targeting a single isolated variant.

Confounding Dependencies. Although sampling and value dependencies constitute distinct mechanisms, they frequently confound one another in practice. For example, within-sibship analyses intentionally leverage sibling cohorts to isolate genetic from environmental effects, thereby introducing both structural sampling dependencies and inherent genetic value dependencies simultaneously.

Dataset Curation. We obtained variants from the ClinVar database (GRCh37/hg19 build) and applied a rigorous filtering pipeline: (1) minimum review status of one star (criteria provided by submitter), excluding variants lacking assertion criteria; (2) deduplication of variants at identical genomic coordinates (chr:pos:ref¿alt); (3) group size constraints, retaining genes containing 5–50 variants after filtering; and (4) class balancing to ensure equal representation of pathogenic and benign variants within each gene group. For the sampling dependency experiment, we used a clinical HCM gene panel; after applying group size filters, 205 gene groups remained. Sequence contexts of 512 base pairs centered on each variant were extracted from the hg19 reference genome.

Group-LiRA Training Setup. For the Group-LiRA evaluation on the curated ClinVar dataset, we fine-tuned a 500M-parameter Nucleotide Transformer v2 in a head-only configuration to predict variant pathogenicity. We trained exactly 64 shadow models to adequately aggregate and compute membership signals across all variants within a targeted gene group.

A.11 Data Curation and Structural Fragmentation

Depending on the underlying genetic architecture, single nucleotide variants (SNVs) and their corresponding phenotypic labels can be structured into distinct learning instances. In monogenic contexts, a single high-penetrance SNV maps directly to a discrete clinical label, such as rare pathogenic variants in MYH7 (hypertrophic cardiomyopathy) or TP53 (Li-Fraumeni syndrome), or the HBB rs334 variant driving sickle cell disease (Zschocke et al., 2023). Conversely, polygenic architectures aggregate hundreds to thousands of low-effect variants to predict continuous phenotypes or composite disease risks, such as endothelial cell polygenic risk scores (EC-PRS) for coronary artery disease (Marston et al., 2025). Furthermore, pleiotropic relationships arise when a single variant influences multiple distinct traits across biological systems, as exemplified by FBN1 mutations affecting the skeletal, ocular, and cardiovascular systems in Marfan syndrome (Sakai et al., 2016).

Depending on the training and tokenization pipeline, an individual's paired haplotypes may be fused into a single unphased data record, modelled as separate phased sequences, or filtered to retain a single representative haplotype (Li et al., 2024). Furthermore, hardware memory constraints and the quadratic scaling of standard self-attention mechanisms preclude genomic language models (gLMs) from processing a complete human genome concurrently to predict comprehensive phenotypic outcomes (Consens et al., 2025). Consequently, an individual's N unique SNV alleles are often partitioned across M localized sequence windows or context-specific representation tasks. This structural fragmentation represents a fundamental departure from conventional tabular data layouts, where an individual's clinical records are preserved as a single, indivisible entity.

A.12 More related works

A.12.1 Attacks on genomics data

Many different types of genomics-related data have been used for genomic studies, for example, RNA expression data (counts of expressions for each gene) (Fang et al., 2024), genotype and phenotype data (Uhler et al., 2013; Fredrikson et al., 2014; Ghosal et al., 2025), genotype-only (Yilmaz et al., 2021; Almadhoun et al., 2020).

Privacy leakage from raw data sharing. It is well-known that raw genomics sequences are highly sensitive; hence, studies usually do not share the raw sequences by default. (Thomas et al., 2024) and (Harmanci & Gerstein, 2016) have shown that direct release of forms of genomic data other than raw sequences (e.g., DNA methylation data, expression data, and even phenotype data) can also lead to re-identification, since the raw genotype can still be inferred from these data types and can be further cross-linked to re-identify an individual.

Privacy leakage from aggregate data sharing. Previous works have also demonstrated that, instead of sharing the raw data, sharing aggregate information about multiple people can lead to privacy leakage: for example, the adversary can leverage the presence or absence of an allele in a cohort (returned by DNA BEACON) to achieve re-identification (Shringarpure & Bustamante, 2015). More powerful attacks are possible by leveraging the (von Thenen et al., 2019; Ayoz et al., 2020) correlations between SNV locations, e.g., linkage disequilibrium (LD), to reduce the number of queries to the database or to group and attribute the multiple people's genotypes back to an individual. When an aggregate query results contain a target and their family, the attack success to find out the target's genotype increases (Almadhoun et al., 2020). (Lan et al., 2025; Paige et al., 2020; Fredrikson et al., 2014) shows that releasing the effect sizes of a genotype to phenotypes or an ML model can be used to extract the genotype of an individual. (Yeom et al., 2018) further showed that the advantage of such model-inversion attacks is bounded by, and reducible to, membership inference advantage, which motivates our train/non-train gap definition of leakage. In addition to genotype reconstruction, membership inference attack is also possible when an aggregate statistics is released, e.g., the minor allel frequency (Homer et al., 2008), classification of generative ML models trained on different types of genomics data (Fang et al., 2024; Oprisanu et al., 2021).

Recent genomic studies leverages genomic language models to train and predict on genomic sequences instead of the conventional tabular representation of the genotypes. (Nemecek et al., 2026) confirmed that genomics language models can memorize an inserted privacy canary. They also demonstrated that when the members and non-members data points are segments of the reference genome, standard MIA is possible. However, in genomics, information like the reference genome is usually treated as a

public information; Memorization of such information does not directly lead to privacy leakage. In addition, different segments of the same organism is usually semantically treated as one private unit, yet the MIA implicitly treated each genomic segment of the whole genome as the private unit. In other words, they did not define what privacy is.

A.12.2 EXISTING PRIVACY-PRESERVING MECHANISMS/ATTEMPTS FOR GENOMICS DATA

Threat models explored by previous works The dependency structure of the private units must be identified and handled to make sure adequate protection is implemented and/or the correct privacy protection is claimed. Previous works have explored local DP (Yilmaz et al., 2021) (each individual applies privacy-preserving mechanisms locally to their data before sending to the data collector), distributed DP (Fang et al., 2024) (all participants collectively apply the privacy-preserving mechanisms) or central DP (Almadhoun et al., 2020; Uhler et al., 2013; Lange, 2024) (the central data curator applies the privacy-preserving mechanisms). In this work, we only consider central DP where the trusted data curator is an entity like a hospital that conducts the genomics study and release the results to external parties later.

A.12.3 INTRA-PERSONAL CORRELATION

Other than inter-personal dependency, there is also dependency within one individual (intra-personal): correlations between single nucleotide variants (SNVs) and correlations between SNVs to the labels. Genomic variants do not occur independently; certain sets of SNVs are rarely separated by recombination events, leading to strong evolutionary correlations known as linkage disequilibrium (LD). Furthermore, strong biological correlations exist between variants located within shared structural or functional units. For instance, SNVs within the same codon are highly correlated because they jointly determine the outcome/labels (e.g., the translated amino acid) and are subject to shared evolutionary selection pressures. Similarly, variants within the same gene or functional region (e.g., promoters or enhancers) often exhibit co-variation or epistatic interactions due to their combined impact on specific biological pathways. Finally, beyond these biological phenomena, there are also non-causal, purely statistical correlations that can arise as artifacts of specific dataset curation and sampling biases.

A.12.4 DEPENDENCY HANDLING

The dependency structure of the private units must be identified and handled to make sure adequate protection is implemented and/or the correct privacy protection is claimed. Group-DP and client-DP can be applied for arbitrary dependency structure within the dataset, providing protection against the worst-case bound assuming the elements in the group or client being perfectly correlated. This is usually not the case. Some works seek to define dependent DP definitions by leveraging the specific dependency structure of the dataset to get tighter bounds (Liu et al., 2016; Roth & Avella-Medina, 2025):

Dependent DP: (Liu et al., 2016) For any datasets $X(L, \mathcal{G}), X'(L, \mathcal{G}) \in \mathscr{X}$ such that changing one data record from X causes atmost $L-1$ other data records to change following the dependency structure $\mathcal{G}$, an algorithm M is ϵ-dependent DP:

$$\Pr[M(X(L, \mathcal{G})) \in S] \leq e^{\epsilon} \Pr\left[M\left(X'(L, \mathcal{G})\right) \in S\right]$$

Dependent DP usually involves reasoning about the maximum change of the query's output (i.e., sensitivity) Δf of changing one data record (and the subsequent change of $L-1$ data records) in accordance to their correlation structure $\mathcal{G}$. For certain query types (e.g., computing the mean), the global sensitivity of the query is usually clipped to bound (Roth & Avella-Medina, 2025).

Pufferfish privacy (Kifer & Machanavajjhala, 2014) and Bayesian Differential Privacy (BDP) (Yang et al., 2015) Pufferfish is a general privacy framework that extends beyond standard DP. It is not a single algorithm but a general framework to define arbitrary secret pairs and adversary's prior knowledge about the dataset. An algorithm M is said to be $\epsilon-$ Pufferfish private with respect to (S, Q, Θ), where S is the set of secrets we need to hide, Q is the secret pairs the adversary shall not distinguish, and Θ is the distribution class for the dataset (which is the prior knowledge to the adversary), for all $(s_i, s_j) \in Q$ and $\theta \in \Theta$,

$$\Pr[M(X) \in S|s_i, X \sim \theta] \leq e^{\epsilon} \Pr\left[M\left(X\right) \in S|\, s_j, X \sim \theta\right].$$

Bayesian DP is an instantiation of Pufferfish privacy framework, which is a privacy definition proposed to account for the dependency structure within the data points other than dependent DP or group DP.

If a dataset $X = \{x_t, x_{\mathcal{K}}, x_{\mathcal{U}}\}$ where x_t is the target the adversary would like to extract, $x_{\mathcal{K}}$ is the data records known by the adversary in advanced, and $x_{\mathcal{U}}$ is the data records not known by the adversary, neighbouring dataset $X' = \{x'_t, x_{\mathcal{K}}, x_{\mathcal{U}}\}$, $x_{-t} = X \setminus \{x_t\}$, $x'_{-t} = X' \setminus \{x_t\}$, the dependency structure between the data points are known to everyone including the adversary, Bayesian DP leakage (BDPL) with respect to the adversary $\mathcal{A}$ for algorithm M is defined as

$$BDPL_{\mathcal{A}} = \sup_{x_t, x'_t, x_{\mathcal{K}}, S} \log \frac{\Pr[M(X) \in S|x_t, x_{\mathcal{K}}]}{\Pr\left[M\left(X'\right) \in S|\, x'_t, x_{\mathcal{K}}\right]} \leq \epsilon,$$

where $\Pr[M(X) \in S|x_t, x_{\mathcal{K}}] = \sum_{x_{\mathcal{U}}} \Pr[M(X) \in S|X] \Pr[x_{\mathcal{U}}|x_t, x_{\mathcal{K}}]$.

The algorithm M satisfy ϵ-BDP if

$$\sup_{\mathcal{A}} BDPL_{\mathcal{A}} \leq \epsilon.$$

BDPL can be written as

$$BDPL_{\mathcal{A}} = \sup(\log \frac{\Pr[x_t|M(X) \in S, x_{\mathcal{K}}]}{\Pr\left[x'_t|\, M\left(X'\right) \in S, x_{\mathcal{K}}\right]} - \log \frac{\Pr[x_t|x_{\mathcal{K}}]}{\Pr\left[x'_t|\, x_{\mathcal{K}}\right]}),$$

and BDP bounds the adversary's gain on the belief of the target x_t after releasing the output $M(X)$ over the prior belief on x_t.

Equivalence of DP and BDP: Note that standard DP definition for an ϵ-DP algorithm M can be rewritten as $\sup_{t, x_t, x'_t, x_{-t}, S} \log \frac{\Pr[M(X) \in S|x_t, x_{-t}]}{\Pr[M(X') \in S|x'_t, x_{-t}]} \leq \epsilon$. That says, when *all data records are independent* or *all possible adversaries $\mathcal{A}$ know the entire dataset other than the target data record* (i.e., $x_{\mathcal{U}} = \emptyset, \forall \mathcal{A}$), ϵ-DP is equivalent to ϵ-BDP.

Privacy Granularity Privacy granularity defines the smallest unit of information to reason about the privacy leakage.

Person-level granularity treats all features from one person as the unit of private information; we reason the privacy leakage of one person as a whole. It is widely used in genomics studies (Uhler et al., 2013; Lange, 2024). When person-level privacy is considered, the value dependency among multiple people (e.g., family members) must be analyzed. (Almadhoun et al., 2020) proposed a family-structure aware adaptation of standard DP to consider **family-level** privacy by logarithmically scaling the sensitivity with the family size[1]. (Lange, 2024) also proved that for a counting query of Mendelian correlated pedigree, an $\epsilon-$DP algorithm with at most m randomly correlated records is no better than $(m-1)\epsilon$-Bayesian DP (definition in Section A.12.4).

Attribute-level granularity protects the privacy of each attribute. When such privacy granularity is used, two types of value dependency shall be carefully analyzed: the correlations of attributes within a person (intra-personal) and the correlation across different people (inter-personal). Prior heuristic approaches like VESPA (Al-Saidi et al., 2026) attempt to empirically protect attribute privacy by applying position-specific perturbations directly to token embeddings. However, by ignoring the dependencies between different genomic positions or the kinship structure across people. Such defences would fall into the classic pitfalls of independent attribute protection (medina et al., 2021) or attacks leveraging the kinship structure (Almadhoun et al., 2020). Yilmaz et al. (2021) proposed a local DP framework that considers the intra-personal dependency. It commented on that having family members releasing their SNV values causes additional leakage. Since it is a local DP framework where everyone releases their SNV value without knowing who else is also releasing, the framework does not address the additional privacy leakage caused by potential kinship structure. Also, since the total number of SNVs to be shared is not known in advance, it does not analyze the composition for the total privacy leakage of one person, which is more commonly used in the central DP case we are considering.

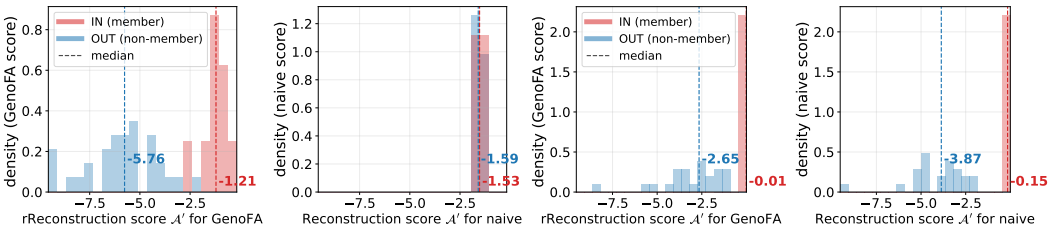
[1]This scaling is empirically determined based on the specific attack, not a theoretical analysis on the global sensitivity.

Training with correlated data Linear mixed models (LMMs) are widely used in genomics to account for relatedness among samples. An LMM extends a fixed-effects regression model by incorporating a random effect whose covariance is parameterized by the genetic relationship matrix (GRM), or kinship matrix, estimated directly from genotype data (Yao & Ochoa, 2023). The GRM captures covariance arising from both recent familial relationships and broader shared ancestry. Unlike PCA-based approaches, which assume that relatedness can be adequately represented in a low-dimensional subspace, LMMs impose no low-rank assumption and can model the complex, high-dimensional correlation structure generated by the many distant relatives present in large human cohorts (Yao & Ochoa, 2023). Furthermore, (Abowd et al., 2013) demonstrates how the sensitivity of LMM parameters can be empirically bounded for a given dataset.

A.13 ADDITIONAL RESULTS: HOW GENOFA IMPACT THE RECONSTRUCTION SCORES

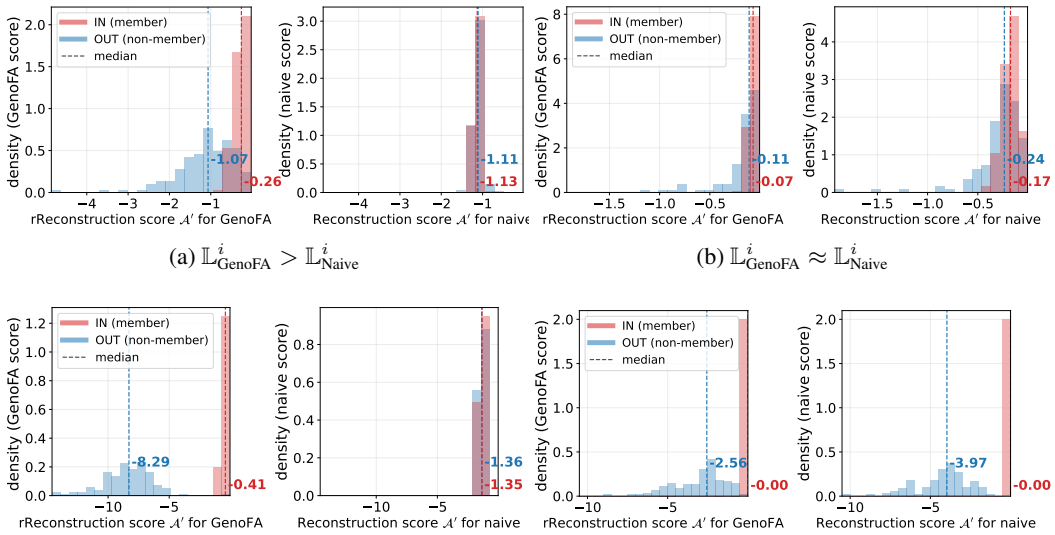
Different cases of $\mathcal{A}'_{\text{in}}(\boldsymbol{x}^i, y^i)$'s and $\mathcal{A}'_{\text{out}}(\boldsymbol{x}^i, y^i)$ are shown in Figure 6. For example, there are cases where GENOFA decreases the validation reconstruction score significantly, this is likely due to the model generalizes poorly onto that family; hence more family members amplifies that bias, making the overall reconstruction less successful for this target person, as shown in Figure 6a. Also, for some person i, the naive reconstruction leakage already reaches 1 (i.e., $\mathbb{L}^i_{\text{Naive}} = 1$). In such cases, GENOFA cannot increase the leakage anymore, as shown in Figure 6b.

Similar to plots Figures 1 and 6 plotted for NT models, we plot the same plots for MLP models in Figure 7.



(a) $\mathbb{L}^i_{\text{GenoFA}} > \mathbb{L}^i_{\text{Naive}}$ but GENOFA decreases ϕ_{out} (b) $\mathbb{L}^i_{\text{GenoFA}} \approx \mathbb{L}^i_{\text{Naive}} \approx 1$ (Both saturate at 1)

Figure 6: Distribution and separation of $\mathcal{A}'_{\text{in}}(\boldsymbol{x}^i, y^i)$'s and $\mathcal{A}'_{\text{out}}(\boldsymbol{x}^i, y^i)$ for FA and naive attacks. 50 shadow models with NT.



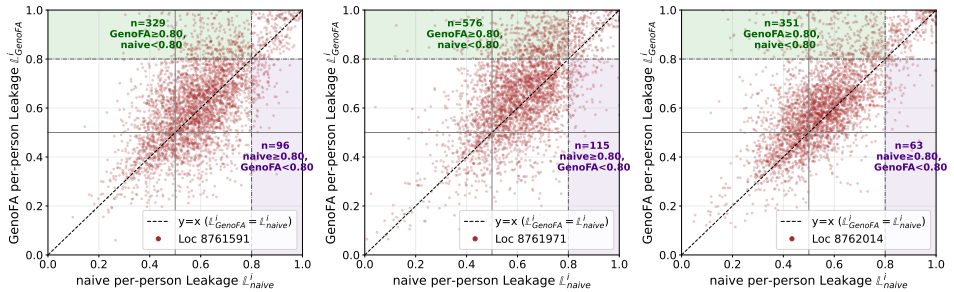
(a) $\mathbb{L}^i_{\text{GenoFA}} > \mathbb{L}^i_{\text{Naive}}$ (b) $\mathbb{L}^i_{\text{GenoFA}} \approx \mathbb{L}^i_{\text{Naive}}$

(c) $\mathbb{L}^i_{\text{GenoFA}} > \mathbb{L}^i_{\text{Naive}}$ but GENOFA decreases ϕ_{out} (d) $\mathbb{L}^i_{\text{GenoFA}} \approx \mathbb{L}^i_{\text{Naive}} \approx 1$ (Both saturates at 1)

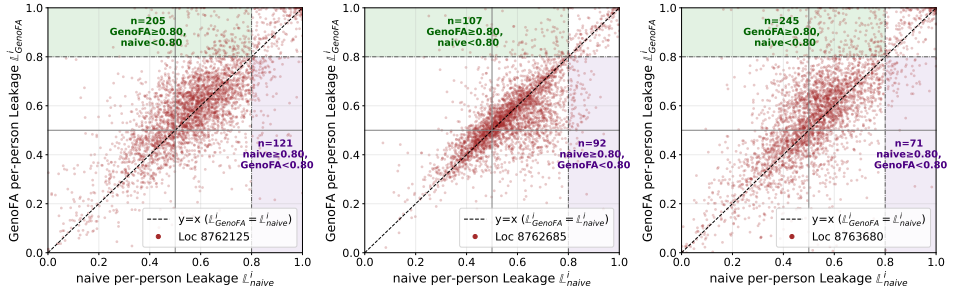
Figure 7: Distribution and separation of $\mathcal{A}'_{\text{in}}(\boldsymbol{x}^i, y^i)$'s and $\mathcal{A}'_{\text{out}}(\boldsymbol{x}^i, y^i)$ for FA and naive attacks. 256 shadow models with MLP.

A.14 ADDITIONAL RESULTS: $\mathbb{L}^i_{\text{GENOFA}}$ VS $\mathbb{L}^i_{\text{NAIVE}}$ FOR MULTIPLE GENOMIC SITES

Figure 8 shows the $\mathbb{L}^i_{\text{GenoFA}}$ vs $\mathbb{L}^i_{\text{Naive}}$ for multiple genomic sites rs53576 (sim) using NT models. Taking Figure 8c as example, which is the same one as Figure 2a: to see an aggregate results, we pick a threshold (e.g., $\tau = 0.8$ in the plots): 351 individuals have only $\mathbb{L}^i_{\text{GenoFA}} \geq \tau$ (green shaded box), which is about $5\times$ the individuals whose $\mathbb{L}^i_{\text{Naive}} \geq \tau$ (purple shaded box), confirming for certain genomic location of attack, GENOFA can cause more individuals to have larger privacy leakage. The numbers of individuals in the green shaded boxes are different, indicating for different sites (attributes) to attack, the number of persons having $\mathbb{L}^i_{\text{GenoFA}} > \mathbb{L}^i_{\text{Naive}}$ is different. Figure 9 shows the fraction of persons with $\mathbb{L}^i_{\text{GenoFA}} \geq \tau \cap \mathbb{L}^i_{\text{Naive}} < \tau$ (fraction of data points in the green shaded box) and $\mathbb{L}^i_{\text{Naive}} \geq \tau \cap \mathbb{L}^i_{\text{GenoFA}} < \tau$ (fraction of data points in the purple box) for different values of τ.

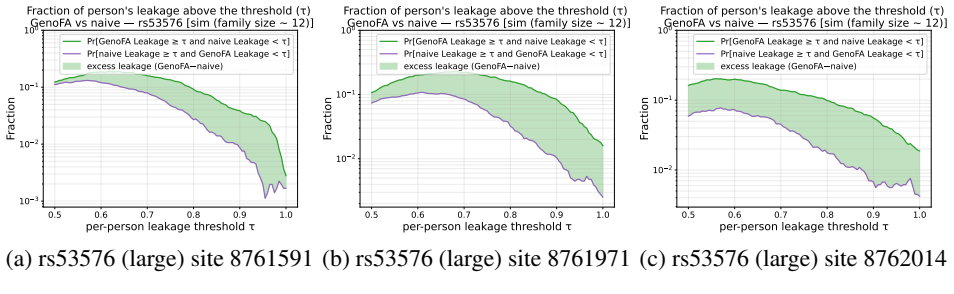


(a) rs53576 (large) site 8761591 (b) rs53576 (large) site 8761971 (c) rs53576 (large) site 8762014

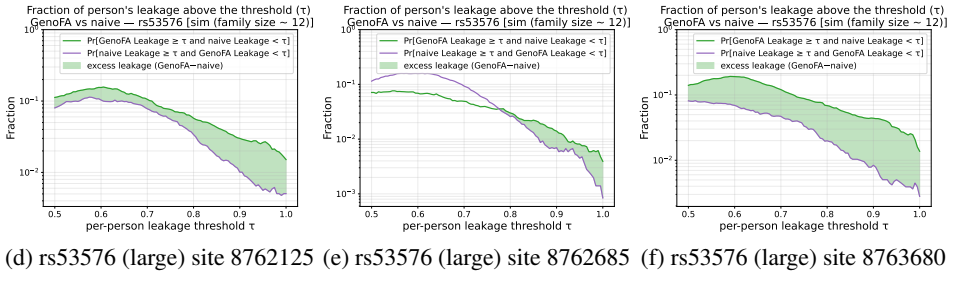


(d) rs53576 (large) site 8762125 (e) rs53576 (large) site 8762685 (f) rs53576 (large) site 8763680

Figure 8: $\mathbb{L}^i_{\text{GenoFA}}$ vs $\mathbb{L}^i_{\text{Naive}}$ for different sites of rs53576 (large) NT model



(a) rs53576 (large) site 8761591 (b) rs53576 (large) site 8761971 (c) rs53576 (large) site 8762014

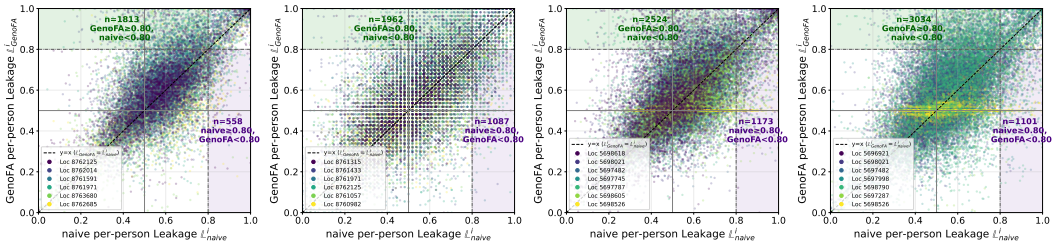


(d) rs53576 (large) site 8762125 (e) rs53576 (large) site 8762685 (f) rs53576 (large) site 8763680

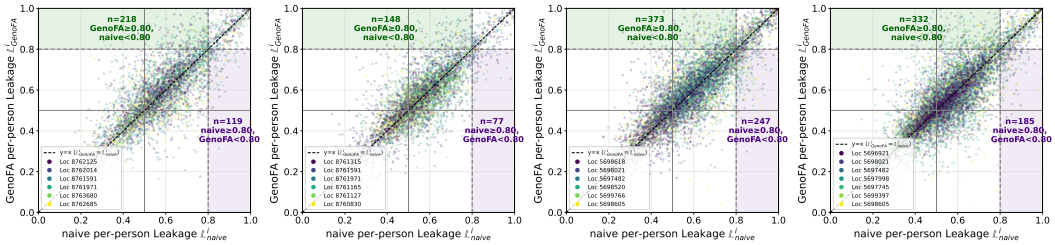
Figure 9: $\mathbb{L}^i_{\text{GenoFA}} \geq \tau \cap \mathbb{L}^i_{\text{Naive}} < \tau$ and $\mathbb{L}^i_{\text{Naive}} \geq \tau \cap \mathbb{L}^i_{\text{GenoFA}} < \tau$ for different sites of rs53576 (sim) NT model

A.15 Additional results: $\mathbb{L}^i_{\text{GenoFA}}$ vs $\mathbb{L}^i_{\text{Naive}}$ for all datasets and models

Figure 10 and 11 summarize the $\mathbb{L}^i_{\text{GenoFA}}$ vs $\mathbb{L}^i_{\text{Naive}}$ for all datasets (real and sim) and models (NT and MLP).

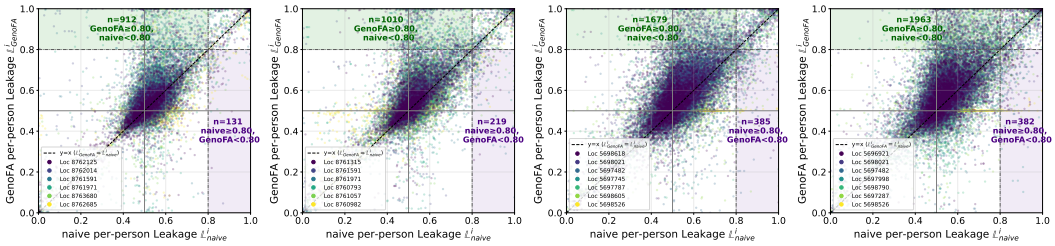


(a) rs53576 (large) (b) rs237890 (large) (c) rs2291844 (large) (d) rs10838549 (large)

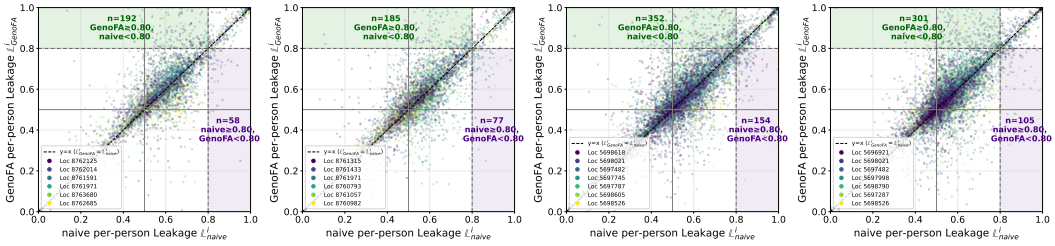


(e) rs53576 (small) (f) rs237890 (small) (g) rs2291844 (small) (h) rs10838549 (small)

Figure 10: $\mathbb{L}^i_{\text{GenoFA}}$ vs $\mathbb{L}^i_{\text{Naive}}$ for all sites for all datasets (NT model)



(a) rs53576 (large) (b) rs237890 (large) (c) rs2291844 (large) (d) rs10838549 (large)

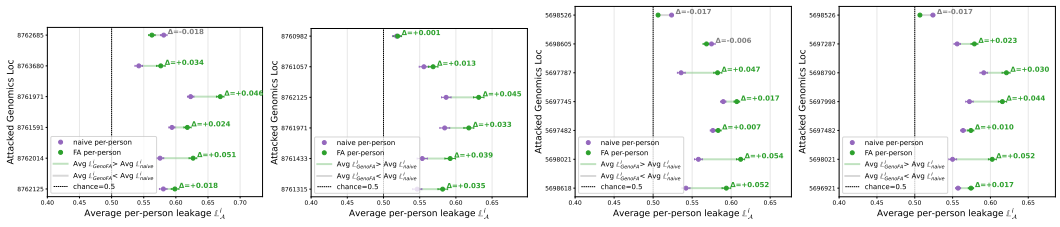


(e) rs53576 (small) (f) rs237890 (small) (g) rs2291844 (small) (h) rs10838549 (small)

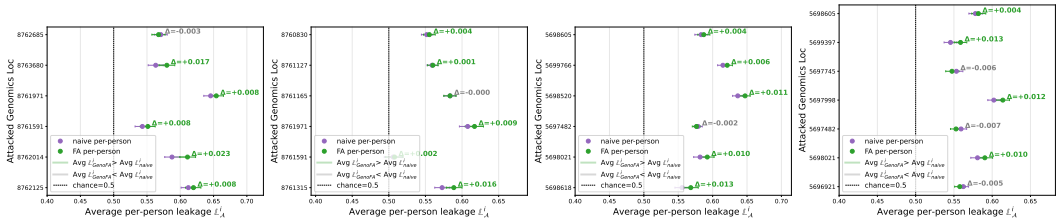
Figure 11: $\mathbb{L}^i_{\text{GenoFA}}$ vs $\mathbb{L}^i_{\text{Naive}}$ for all sites for all datasets (MLP model)

A.16 Additional results: per-person leakage (aggregated over all persons)

Using the same setup as Figure 2b, we plot Figure 12 and 13 to show that FA elevates per-person leakage for all dataset using model NT and MLP respectively.

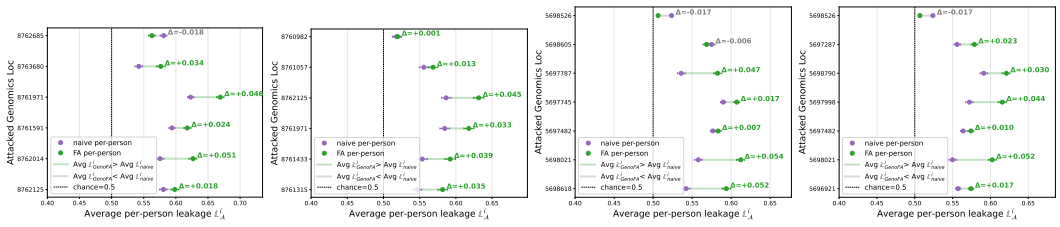


(a) rs53576 (large) (b) rs237890 (large) (c) rs2291844 (large) (d) rs10838549 (large)

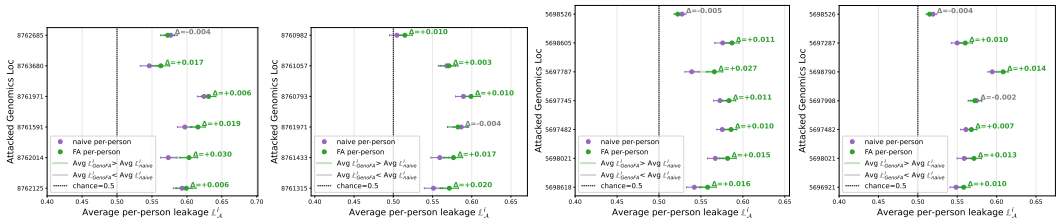


(e) rs53576 (small) (f) rs237890 (small) (g) rs2291844 (small) (h) rs10838549 (small)

Figure 12: Average across all persons of $\mathbb{L}^i_{\text{GenoFA}}$ vs $\mathbb{L}^i_{\text{Naive}}$ for all dataset. NT model.



(a) rs53576 (large) (b) rs237890 (large) (c) rs2291844 (large) (d) rs10838549 (large)

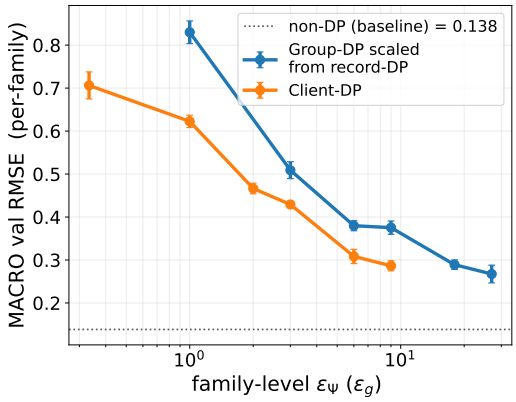


(e) rs53576 (small) (f) rs237890 (small) (g) rs2291844 (small) (h) rs10838549 (small)

Figure 13: Average across all persons of $\mathbb{L}^i_{\text{GenoFA}}$ vs $\mathbb{L}^i_{\text{Naive}}$ for all dataset. MLP model.

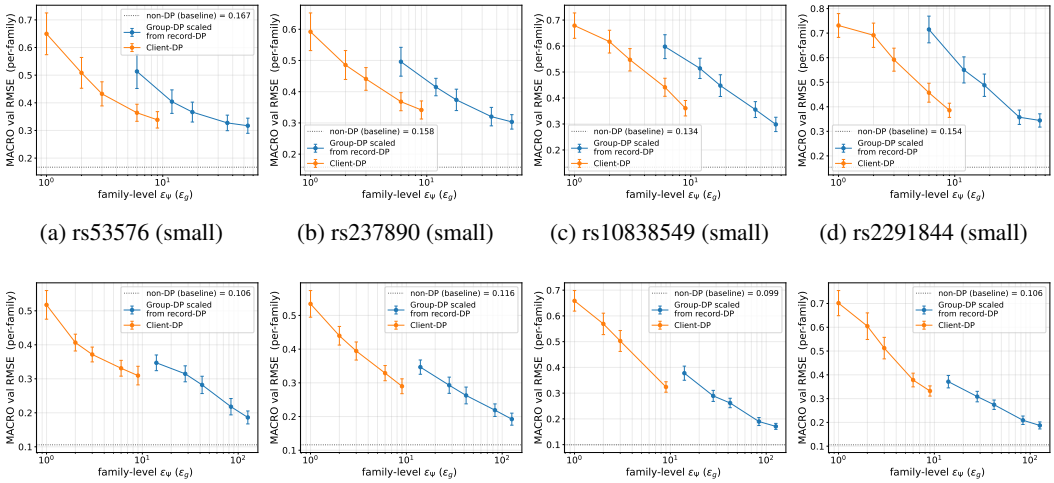
A.17 Additional results: privacy utility tradeoff

Additional privacy-utility tradeoff plots for NT and MLP are shown in Figures 14 and 15



(a) rs237890 (small)

Figure 14: Client-DP can provide better privacy utility tradeoff than group-DP. Model NT.



(a) rs53576 (small) (b) rs237890 (small) (c) rs10838549 (small) (d) rs2291844 (small)

(e) rs53576 (large) (f) rs237890 (large) (g) rs10838549 (large) (h) rs2291844 (large)

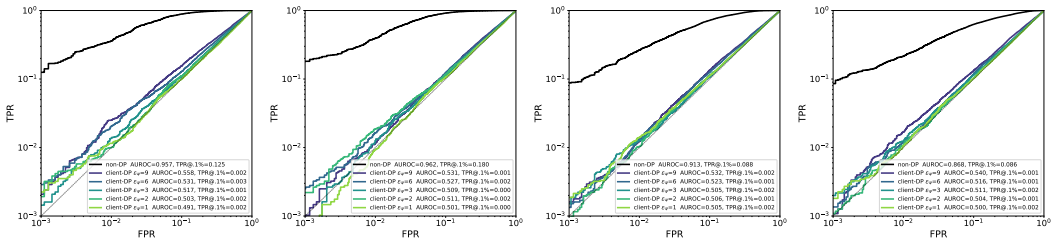
Figure 15: Client-DP can provide better privacy utility tradeoff than group-DP. Model MLP.

A.18 Additional results: Group-LiRA

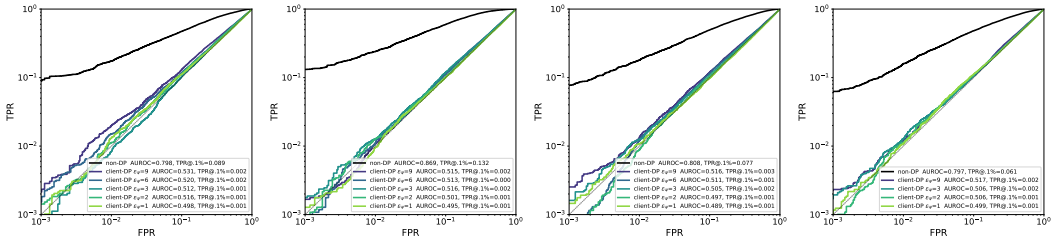
The AUROC and TPR at 0.1% FPR results for Figure 3b is summarized in Table 5. In addition, we show similar plots for different datasets, client-DP and group-DP (scaled from record-DP) in Figure 16 and Figure 17

Table 5: Group-LiRA attack performance results for non-DP and client-DP trained models across various privacy budgets.

Setting	**AUROC**	**TPR @ 0.1%**
Non-DP	0.957	0.125
Client-DP ($\epsilon_\Psi = 9$)	0.558	0.002
Client-DP ($\epsilon_\Psi = 6$)	0.531	0.003
Client-DP ($\epsilon_\Psi = 3$)	0.517	0.001
Client-DP ($\epsilon_\Psi = 2$)	0.503	0.002
Client-DP ($\epsilon_\Psi = 1$)	0.491	0.002

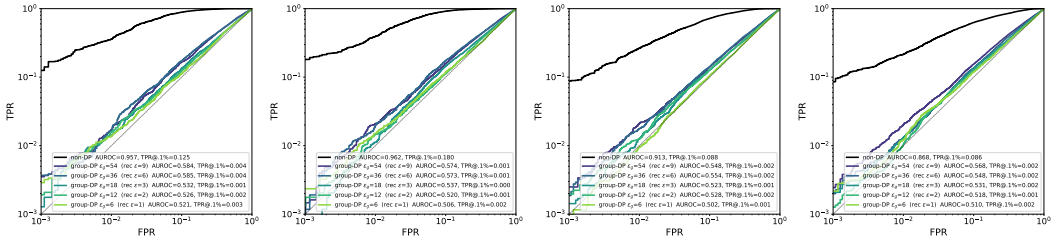


(a) rs53576 (small) (b) rs237890 (small) (c) rs2291844 (small) (d) rs10838549 (small)

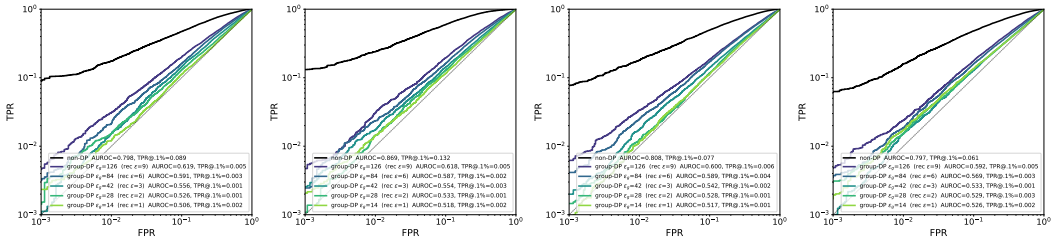


(e) rs53576 (large) (f) rs237890 (large) (g) rs2291844 (large) (h) rs10838549 (large)

Figure 16: Per-family leakage (group-LiRA) (MLP) after client-DP.



(a) rs53576 (small) (b) rs237890 (small) (c) rs2291844 (small) (d) rs10838549 (small)



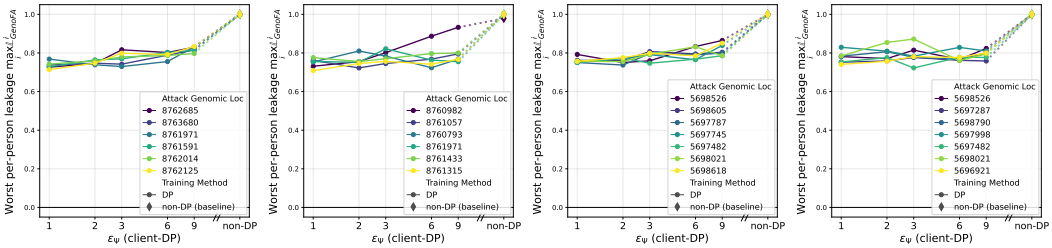
(e) rs53576 (large) (f) rs237890 (large) (g) rs2291844 (large) (h) rs10838549 (large)

Figure 17: Per-family leakage (group-LiRA) (MLP) after record-DP scaled to group-DP.

A.19 Additional results: worst per-person leakage vs privacy budget

we show the worst per-person leakage at various attacked genomic locations for each dataset in Figures 18 and 19.

The values in Figure **??** is summarized in Table 4

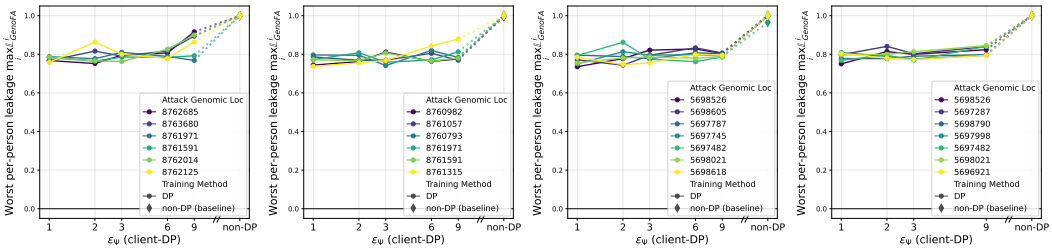


(a) rs53576 (small)

(b) rs237890 (small)

(c) rs2291844 (small)

(d) rs10838549 (small)



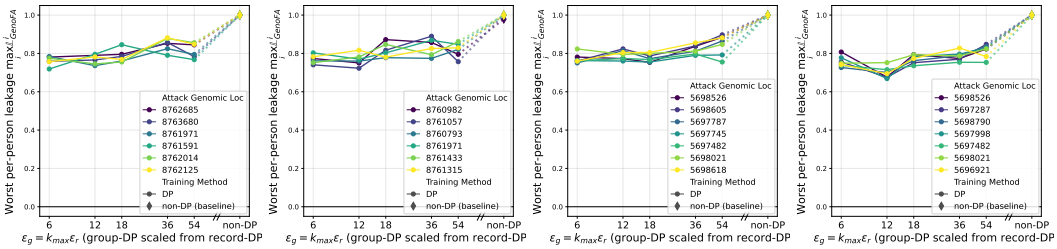
(e) rs53576 (large)

(f) rs237890 (large)

(g) rs2291844 (large)

(h) rs10838549 (large)

Figure 18: Worst per-person leakage GENOFA (MLP) after client-DP

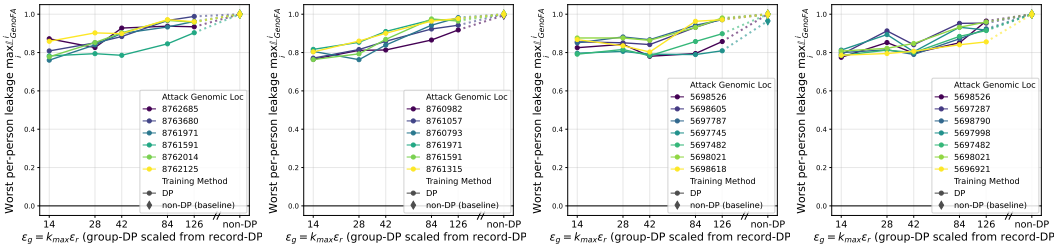


(a) rs53576 (small)

(b) rs237890 (small)

(c) rs2291844 (small)

(d) rs10838549 (small)



(e) rs53576 (large)

(f) rs237890 (large)

(g) rs2291844 (large)

(h) rs10838549 (large)

Figure 19: Worst per-person leakage GENOFA (MLP) after record-DP scaled to group-DP.