PRISM-G: an interpretable privacy scoring method for assessing risk in synthetic human genome data

Alejandro Correa Rojo[1], Yves Moreau[1], Gökhan Ertaylan[2]

[1]ESAT- STADIUS, KU Leuven, Leuven 3001, Belgium

[2]Flemish Institute for Technological Research (VITO), Mol 2400, Belgium

Abstract

The growing use of synthetic genomic data promises broader data access but raises unresolved concerns about privacy risk. We introduce PRISM-G, a model-agnostic framework that summarizes privacy exposure of synthetic genomes across three complementary components: (i) a proximity view that asks whether synthetic individuals lie unusually close to real genomes in genetic-coordinate space; (ii) a kinship view that detects replay of familial or population-structure patterns beyond what is expected by chance; and (iii) a trait-linked view that captures exposure through rare variants and simple membership-inference signals. Each component yields a normalized risk score and a risk-averse aggregation maps these to a 0–100 PRISM-G score. We evaluated PRISM-G on synthetic cohorts generated by a generative adversarial network (GAN), a restricted Boltzmann machine (RBM), and a logic-based SAT-solver (Genomator). Our results show that privacy vulnerabilities concentrate along different axes across models and marker densities, underscoring that a single privacy-based similarity metric is insufficient.

Introduction

The increasing availability of large health data resources, such as biobanks and population cohorts, has accelerated biomedical discovery and clinical research by enabling advanced data analytics and the application of artificial intelligence at population scale (Venturini et al., 2025). With declining DNA

sequencing costs and the rise of population-scale genomics, genomic data has become central to studies of human variation and disease etiology, paving the way toward precision medicine and target therapeutics (Y.-M. Chen et al., 2025; Mani et al., 2025). However, genomes inherently enable identification and are consequently sensitive, making broad data sharing a persistent technological, ethical, and legal challenge due to demonstrated privacy risks and stringent regulatory requirements (Bonomi et al., 2020; Stark et al., 2025). In response, synthetic data generation has emerged as a potential strategy to expand access while mitigating privacy risks and preserving analytical utility without direct reliance on real genomic records (Mendes et al., 2025; Wharrie et al., 2023).

A diverse toolkit now exists for generating synthetic genomic data. Traditional haplotype-based resampling and recombination methods (e.g., HAPNEST) provide scalable, reference-guided genotype synthesis and quality evaluation (Wharrie et al., 2023). More recently, deep generative models, including generative adversarial networks (GANs), restricted Boltzmann machines (RBMs), variational autoencoders (VAEs), and diffusion models, have been adapted to capture linkage disequilibrium, population structure and higher-order dependencies in genomes (Kenneweg et al., 2025; Yelmen et al., 2021, 2023). In parallel, logic-based generators using SAT solvers (e.g. Genomator) cast data synthesis as a constrained satisfaction problem, enforcing biological plausibility and privacy limits by design (Burgess et al., 2024). Collectively, these methods have demonstrated the ability to produce faithful synthetic cohorts with promising applications in population genetics and clinical genomics. As generative modeling continues to mature, synthetic genomic data are poised to become increasingly prominent in genetic research.

Despite the demonstrated feasibility of synthetic data, quantifying privacy risk remains challenging (Arora et al., 2025). In synthetic genomic data, privacy risk relies on similarity-based criteria, such as nearest-neighbor or Hamming-distance proximity between synthetic and real genomes, as proxy for identifiability (Burnard et al., 2025; Yelmen et al., 2023). Beyond distance-based proxies, membership

inference attacks (MIAs) assess if a model or dataset reveals the presence of specific individuals, using shadow generators and a classifier to separate training from holdout participants (Shokri et al., 2017). Although distance metrics have sometimes been taken to imply that synthetic genome data are "privacy-preserving by default," attack-based studies indicate that achieving strong privacy alongside high analytical utility is not uniform across methods. Moreover, models reproducing population-level distributions can still leave individuals susceptible to membership detection (Oprisanu et al., 2022). In practice, personal information can be inferred through data linkage, membership inference, genealogical familial search, DNA phenotyping, and trait inference that leverage external data, population structure, linkage disequilibrium, and functional genomics (Bonomi et al., 2020; Gürsoy et al., 2022). More broadly, there is no consensus or standardized framework for evaluating privacy risks in synthetic data; and, for genomics specifically, there remains a lack of metrics that reflect domain-specific properties while explicitly incorporating the privacy-utility trade-off (Kaabachi et al., 2025).

Genomic data are intrinsically multi-representational: the same individual can be characterized in coordinate space capturing population structure (e.g., eigenanalysis of common variants) (Patterson et al., 2006), in kinship space reflecting familial relatedness and identical-by-descent (IBD) (Manichaikul et al., 2010), and in trait-linked feature space aggregating biologically meaningful signals such as rare-variant burdens (Madsen & Browning, 2009). Each representation has established metrics, for example, population distances and clustering based on principal components analysis (PCA), genetic relationship matrices (GRMs) and IBD statistics, and gene or region burden scores, that can indicate privacy risk when synthetic samples unduly approximate real individuals, recreate close familial ties, or preserve trait-linked outliers.

Building on these representations and metrics, we present **Privacy Risk Integrated Score for Multi-representation Genomes** (PRISM-G), a framework that evaluates synthetic cohorts across three complementary views and applies membership-inference tests based on rare variant content,

calibrating the evidence into a single, interpretable risk score. PRISM-G summarizes privacy exposure via three representation indices: (i) **Proximity leakage (PLI)** asks whether any synthetic genome lies unusually close to a real (holdout) genome in genetic-coordinate space, closer than a normal population structure would explain. (ii) **Kinship replay (KRI)** the synthetic cohort inadvertently recreates family structure or long-range dependence (e.g., close cousins, inflated relatedness, repeated haplotype segments). (iii) **Trait-linked leakage (TLI)** captures exposure that arises when simple membership inference signals or rare-variant burdens or collisions make individuals stand out. Each index yields a unit-interval risk (0-1), combined by a risk-averse "OR-like" aggregator and calibrated to a 0-100 scale using safe and leaky reference generators for interpretability across datasets.

To demonstrate its applicability, we benchmarked three synthetic data generation approaches including GANs, RBMs, and methods based on SAT solvers, and evaluated their privacy risks across genomic representations, using subsets from the 1000 Genomes Project as reference dataset.

Algorithmic basis and background

As previously mentioned, early evaluations of synthetic genomic data often relied on similarity indices, such as distance to the nearest real neighbor, to proxy identity risk. Yet genomic privacy risks are not limited to pointwise closeness: (i) population structure and relatedness create identifiable patterns that enable familial search and relationship inference, and (ii) functional profiles and rare variants can reveal membership or traits even when overall distances look innocuous. Foundational work on PCA provides compact genetic axes that reveal population structure or demographic variation, while GRM and kinship estimation methods operationalize relatedness at scale (Goudet et al., 2018; Patterson et al., 2006). In combination, these approaches can enable re-identification through distant relative matching in large databases. In parallel, privacy studies show MIAs against synthetic data and beacon-

style singling-out driven by rare variants (Oprisanu et al., 2022; Shringarpure & Bustamante, 2015). Taking this into account, PRISM-G's components are designed to operationalize these documented risks using established genomic methodology: coordinates for proximity, GRM/kinship and short-haplotype patterns for relatedness replay, and MIA signals plus rare-variant counts for trait-linked exposure. Table 1 shows an overview of the methodological basis and related work on the methods implemented in PRISM-G.

Component	Algorithmic basis	PRISM-G adaptation	Adapted or based on
PLI	Projects genomes into PCA-derived genetic coordinates and tests whether holdouts are unusually close to synthetics in the lower tail of the distance distribution, benchmarked against holdout-to-holdout distances, with an adversarial nearest-neighbor check to avoid flagging ordinary structure.	Built on PCA for interpretable genetic axes, PRISM-G replaces raw nearest neighbor distance with a calibrated lower-tail closeness test and an adversarial sanity check, yielding a conservative proximity signal.	(Patterson et al., 2006; Yale et al., 2019.)
KRI	Quantifies relatedness using GRM/KING, tests for excess kinship against a safe baseline, scans for short haplotype sharing via short unphased genotype window, and flags spectral anomalies by detecting inflation of the GRM's leading eigenvalue.	PRISM-G combines four baseline-corrected signals: relatedness surplus, kin-edge replay, micro-haplotype collisions, and spectral concentration, into a novel aggregate that detects replayed familial structure and long-range dependence.	(Goudet et al., 2018; Kuo et al., 2023; Manichaikul et al., 2010; Oldoni et al., 2019; Seidman et al., 2020)
TLI	Build on rare-variant burdens to probe membership-inference signals in the synthetic cohort and estimates carrier probabilities from allele frequencies under Hardy-Weinberg equilibrium to flag excess co-ocurrence of rare variants.	PRISM-G implements a simple nearest neighbor distance built from burden scores, and a rare-variant check that highlights unusual co-ocurrence of uncommon variants. Together the results are mapped into a single score.	(Madsen & Browning, 2009; Shokri et al., 2017; Shringarpure & Bustamante, 2015)

Table 1. Overview of the methodological foundations and related work for the PRISM-G component metrics.

Materials and methods

Model description

Given a generative model, let R_{tr} denote the real training cohort of m individuals and n single nucleotide polymorphisms (SNPs) used to produce a synthetic cohort S, and let R_{ho} denote the real holdout cohort reserved for evaluation. PRISM-G evaluates S against R_{ho} in population coordinate space, kinship space, and trait linked feature space, respectively. We first describe the three components used to estimate PRISM-G. Each component produces a score in [0, 1], where 0 indicates no detected leakage on that axis and 1 indicates maximal exposure.

Proximity leakage index

The PLI metric quantifies whether synthetic samples lie unusually close to real holdout individuals in genetic coordinate space, beyond what would be expected from natural clustering among real individuals. Genotypes are embedded into a d dimensional space (for example, principal components computed after standardizing genotypes using allele frequencies estimated from R_{tr}) (Patterson et al., 2006). Let $X_{\mathrm{ho}} \in \mathbb{R}^{n_{\mathrm{ho}} \times d}$ and $X_S \in \mathbb{R}^{n_S \times d}$ be the embedding for R_{ho} and S. For each holdout individual i, define the nearest synthetic distance:

$$d_i^{\mathrm{syn}} = \min_j \left\| x_i^{\mathrm{ho}} - x_j^S \right\|_2 \tag{1}$$

and the baseline nearest real distance to other holdout samples:

$$d_i^{\text{real}} = \min_{j \neq i} \left\| x_i^{\text{ho}} - x_j^{\text{ho}} \right\|_2 \tag{2}$$

We compare the lower tails of the distance distributions estimated from equations 1 and 2 via a quantile ratio:

$$\rho_q = \frac{Q_q\left(\{d_i^{\text{syn}}\}\right)}{Q_q\left(\{d_i^{\text{real}}\}\right)} \tag{3}$$

and define a proximity signal:

$$r_\rho = \text{clamp}\left(1 - \rho_q, 0{,}1\right) \tag{4}$$

where $\text{clamp}(x, 0{,}1) = \min\{1, \max\{0, x\}\}$.

To add an adversarial check, we ask whether a holdout point is typically closer to another holdout than to any synthetic sample (Yale et al., 2019). Let $a_i = 1[d_i^{\text{real}} < d_i^{\text{syn}}]$ and $\text{A} = \frac{1}{n_{\text{ho}}} \sum_i a_i$; map this to risk by $r_{\text{A}} = \max\left\{0, \frac{0.5 - \text{A}}{0.5}\right\}$. The PLI is calculated as:

$$\text{PLI} = \max\left(r_\rho, r_{\text{A}}\right) \tag{5}$$

Kinship replay index

The KRI metric measures whether the synthetic cohort recreates family structure or long-range dependence beyond safe expectations. We compute GRMs for R_{ho} and S after standardizing genotypes

by training set allele frequencies, yielding K^{ho} and K^{S} (Manichaikul et al., 2010). KRI aggregates four signals, each baseline corrected against a safe or bootstrap reference.

First, the **replay** component (r_{replay}) asks whether close kin observed in R_{ho} are echoed in S. For kinship thresholds $\theta \in \{0.10, 0.125, 0.25\}$ we measure the fraction M of real close-kin pairs that have a synthetic pair within a small tolerance $\pm\,\tau$ of the same kinship; we subtract a null rate M_0 obtained by randomly permuting rows and columns of K^{S}, and set:

$$r_{\mathrm{replay}} = \mathrm{clamp}\left(\frac{M - M_0}{1 - M_0}, 0,1\right) \tag{6}$$

Second, the **internal kinship excess** (r_{IKE}) tests for an overall surplus of relatedness in S. Let $f(\theta) = \Pr[K_{ij}^{S} \geq \theta, i < j]$. Using a safe size and MAF matched baseline $f_0(\theta)$ (or a holdout bootstrap when a safe set is unavailable), define:

$$r_{\mathrm{IKE}} = \mathrm{clamp}\left(\frac{f(\theta) - f_0(\theta)}{1 - f_0(\theta)}, 0,1\right) \tag{7}$$

Third, the **micro haplotype collision** (r_{HAP}) searches for excess reuse of short genotype patterns without requiring phase (Seidman et al., 2020). We slide fixed width windows of k SNPs and encode per window genotype patterns; for each pair of synthetic individuals, we count shared windows and compute a mean collision rate c. After subtracting a matched baseline c_0, we set:

$$r_{\mathrm{HAP}} = \mathrm{clamp}\left(\frac{c - c_0}{c_{\max} - c_0}, 0,1\right) \tag{8}$$

with $c_{\max}$ taken as high percentile to stabilize scaling.

Fourth, the **spectral anomaly** (r_{SPEC}) capture concentration of relatedness via the top eigenvalue of K^S. Let $s = \lambda_1/\mathrm{trace}(K^S)$ and s_0 be its baseline expectation; define:

$$r_{\mathrm{SPEC}} = \mathrm{clamp}\left(\frac{s - s_0}{s_{\max} - s_0}, 0,1\right) \tag{9}$$

The KRI metric is estimated as:

$$\mathrm{KRI} = \max(r_{\mathrm{replay}}, r_{\mathrm{IKE}}, r_{\mathrm{HAP}}, r_{\mathrm{SPEC}}) \tag{10}$$

Trait-linked leakage index

The TLI metric captures exposure through biologically meaningful features, focusing on membership inference and on the preservation of rare or distinctive signals. Let c_i denote trait features for individual i (rare variant component burdens) (Madsen & Browning, 2009).

For **membership inference advantage**, to probe whether training presence leaves a detectable trace in the synthetic cohort S, we construct a reference summary features from c_i:

$$d_S(i) = \min_{s \in S} \|c_i - c_s\| \tag{11}$$

Using the S-referenced features, we score the training, and holdout sets with the negative distances $s(i) = -d_S(i)$ as membership scores, since being closer to S implies stronger membership evidence. We label the train individuals as 1 and holdout as 0, evaluate the area under the receiver operating characteristic (ROC) curve A, and define:

$$r_{\text{mia}} = \text{clamp}(2\text{A} - 1, 0,1) \tag{12}$$

so random performance maps to zero and perfect separation maps to 1.

For **uniqueness and rarity match**, we evaluate whether rare signals occur more often in S than expected under a null driven by training allele frequencies (Shringarpure & Bustamante, 2015). Define the rare set $R = \{j : MAF_j^{\text{train}} < \tau\}$. For are SNP j with training allele frequency p_j, under Hardy-Weinberg equilibrium approximation, the diploid carrier probability is $q_j \approx 2p_j$, and the null probability in a cohort of size n_s that at least two synthetic individuals carry the variant is:

$$\Pr[X_j \geq 2] = 1 - (1 - q_j)^{nS} - n_S q_j (1 - q_j)^{n_S - 1} \tag{13}$$

Let U be the observed fraction of rare sites with at least two carriers in S and U_0 the average of the null probabilities above; set:

$$r_{\text{uniq}} = \text{clamp}\left(\frac{U - U_0}{1 - U_0} - 1, 0,1\right) \tag{14}$$

The TLI metric is estimated as:

$$\text{TLI} = \max(r_{\text{mia}}, r_{\text{uniq}}) \tag{15}$$

Aggregation of component risks

Each component produces a unit-interval risk score with higher values indicating greater privacy risk: PLI for coordinate proximity, KRI for kinship replay, and TLI for trait linked exposure. To obtain a single cohort-level risk, we combine component scores using a risk-averse "OR-like" aggregator that preserves monotonicity and ensures that a single high-risk component cannot be masked by lower values in the others. Specifically, let $r_{\mathrm{PLI}}, r_{\mathrm{KRI}}, r_{\mathrm{TLI}} \in [0, 1]$ denote the component scores and $w_{\mathrm{PLI}}, w_{\mathrm{KRI}}, w_{\mathrm{TLI}} \geq 0$ weights that sum to one. The aggregated raw risk is defined as:

$$R_{\mathrm{raw}} = 1 - \prod_{m \in \{\mathrm{PLI},\, \mathrm{KRI},\, \mathrm{TLI}\}} (1 - w_m r_m) \tag{16}$$

This construction has three practical advantages. First, it is **monotone** in every component: increasing any r_m increases R_{raw}. Second, it is **risk-averse**: if any component approaches one, R_{raw} also approaches one regardless of the others. Third, it is **scale-free** with respect to cohort size, since components are already normalized to [0, 1].

When a sub-diagnostic inside KRI is not applicable (for example, the replay term when the holdout contains no close relatives), that sub-term is omitted from KRI's internal maximum without renormalizing other components; this preserves comparability across datasets.

Calibration and estimation of PRISM-G score

Raw aggregation yields a number in $[0, 1]$ that is not directly interpretable across datasets. We therefore calibrate R_{raw} using two reference generators that anchor the lower and upper ends of the scale:

1. a **safe baseline** that preserves allele frequencies but intentionally removes dependence structure (for example, an allele-frequency–matched marginal sampler), and
2. one or more **leaky baselines** that intentionally overfit structure (for example, copycat, kinship-preserving variants or data itself).

For each baseline we compute the distribution of R_{raw}. Let α denote the calibrated lower anchor, taken as the median (or 95th percentile, for stricter calibration) of R_{raw} under the safe baseline, and let β denote the upper anchor, taken as the median (or 5th percentile) under the leaky baseline. We then map the observed raw risk for a model of interest to a standardized 0-100 scale:

$$\text{PRISM-G} = 100 \times \text{clamp}\left(\frac{R_{\text{raw}} - \alpha}{\beta - \alpha}, 0, 1\right) \tag{17}$$

This linear calibration is monotone, model-agnostic, and yields scores that are comparable across datasets if the same anchors are used.

For interpretability, we group scores into qualitative bands: **Green**, **Amber**, and **Red**, indicating whether a method yields a safe, leaky, or risky dataset. These bands can be tuned to application-specific tolerances by shifting cutpoints or by defining them from anchor distributions (for example, set the Green/Amber boundary at the 95th percentile of the safe baseline and the Amber/Red boundary at the 5th percentile of the leaky baseline).

Evaluation

Generative models

We evaluated PRISM-G on synthetic cohorts generated from three model families: (i) **GANs**, (ii) **RBMs** as introduced by Yelmen et al., and (iii) **Genomator**, a logic-based generator that frames synthesis as a satisfiability (SAT) problem. Briefly, GANs pair a generator that maps random noise to genotype vectors with a discriminator that distinguishes real from synthetic samples: adversarial training drives the generator to match allele frequencies, population structure, and pairwise correlations, and conditional or convolutional variants provide control over ancestry and scalability. RBMs learn the joint distribution of genotype vectors via an energy function; after training (e.g., with contrastive divergence) they generate new genomes by alternating Gibbs updates, capturing linkage disequilibrium and higher order dependencies in the data (Yelmen et al., 2023). In contrast, Genomator encodes admissible genotype configurations as logical constraints and uses SAT solving to sample synthetic genomes that adhere to specified biological and privacy rules, with a tunable parameter controlling distances to training data and within the synthetic cohort (Burgess et al., 2024).

As reference, we used the 1000 Genomes Project Phase 3 (1KGP) panel comprising 2,504 individuals. This served as the source for training/holdout partitions and for aligning markers across experiments.

Synthetic datasets

For the GAN and RBM baselines, we used the synthetic cohorts released by Yelmen et al. (2023), generated from 1KGP. For Genomator, we generated new synthetic cohorts from the same training splits under three Hamming distance settings $H \in \{1, 10, 50\}$, which regulate proximity between training and synthetic samples and among synthetic samples.

All analyses were performed on two datasets: one from chromosome 15 with 10,000 SNPs and one from chromosome 1 with 65,535 SNPs. For each dataset and generator, the synthetic cohort size matched the corresponding holdout cohort to enable privacy evaluation with PRISM G (PLI, KRI, TLI.

Parameter settings

For PRISM-G, each dataset was split into training (80%) and holdout (20%) sets. The training set was used to fit baselines for the component metrics and to define embeddings; the holdout set was used for final estimation of the component metrics and for membership-inference evaluation.

For PLI metrics, we computed genetic coordinates using principal component analysis (PCA) on the training set and retained the first 10 components for all analyses. PLI used a fixed lower-tail quantile of 0.01 in the distance quantile ratio to focus on the closest real–synthetic pairs.

For KRI metrics, we used a kinship replay threshold of 0.125 (approximately first-cousin relatedness) with a tolerance $\tau = 0.01$ when matching close-kin edges in the real data to those observed in the synthetic cohort. The r_{IKE} metric was evaluated at thresholds $\theta \in \{0.10, 0.125, 0.25\}$ and KRI took the maximum excess across these levels.

For TLI metrics, rare variants were defined by minor-allele frequency (MAF) < 0.001, with allele frequencies estimated from the training set. For membership-inference risk, we used a simple nearest-neighbor exposure test with the estimated rare-burden features: we fitted a 1-NN search over the synthetic coordinates using Manhattan distance, computed each real individual's distance to its nearest synthetic neighbor, and negated those distances so that smaller distance represents higher exposure. We labeled real individuals as 1 = training and 0 = holdout, then computed the ROC area under the curve (AUC) of the exposure scores; the AUC was linearly mapped to a unit-interval risk as specified in the model description. For the estimation of raw values, we used equal weights for the component metrics.

Calibration anchors

We used two references to calibrate PRISM-G. The safe dataset is a sampler from a binomial distribution: for each SNP j with training-set allele frequency p_j^{train}, and for each synthetic individual i, we draw an independent diploid genotype:

$$X_{ij} \sim \text{Binomial}\left(2, p_j^{\text{train}}\right), \quad X_{ij} \in \{0,1,2\} \tag{17}$$

so that

$$\Pr\left[X_{ij} = g\right] = \binom{2}{g}\left(p_j^{\text{train}}\right)^g\left(1 - p_j^{\text{train}}\right)^{2-g} \quad (g = 0, 1, 2) \tag{18}$$

Draws are independent and identically distributed (i.i.d.) across i and j, using the same markers and sample size as the holdout set. This construction preserves site-level frequencies $\mathbb{E}[X_{ij}/2] = p_j^{\text{train}}$ but removes linkage disequilibrium, kinship, and trait-linked correlations.

The leaky dataset is a copycat, kin-doped reference that intentionally overpreserves structure: each synthetic genome X_{ij} is templated to a training individual $t(i)$ with small per-site perturbation (e.g., Hamming distance $H\left(X_i, X_{t(i)}\right) \leq \delta$), and a subset of templates is duplicated or paired to induce close-kin edges (target kinship θ in the range 0.25 – 0.125). This produces synthetic cohorts unusually close to training genomes and enriched for relatedness, serving as a high-risk anchor on the PRISM-G scale.

Results

Overview of PRISM-G components

We first evaluated the three genomic data generators using the defined submetrics before computing PLI, KRI, and TLI, with 1KGP as the training dataset. Figures 1 and 2 report the submetrics for the two marker panels. For the 10,000 SNPs panel from chromosome 15 (Fig. 1a; Supplementary Table 1), PLI submetrics showed that GAN and RBM have lower-tail quantile ratios above one $(1.06 < \rho_q < 1.10)$, yielding $r_\rho = 0$; thus the quantile term contributes no risk, whereas the adversarial proximity rises modestly above chance $(0.32 < r_{\mathrm{A}} < 0.35)$. Genomator exhibited a tunable proximity pattern: at $H = 1$ the synthetic set lay unusually close to holdout individuals $(\rho_q = 0.19, r_\rho = 0.81, r_{\mathrm{A}} = 0.82)$, and proximity attenuated as H increased $(H = 10: \rho_q = 0.41, r_\rho = 0.59, r_{\mathrm{A}} = 0.80;\ H = 50: \rho_q = 0.74, r_\rho = 0.26, r_{\mathrm{A}} = 0.73)$. KRI submetrics (Fig. 1b, Supplementary Table 2) were driven primarily by the spectral term for Genomator $(0.70 < r_{\mathrm{SPEC}} < 0.78)$ with small contributions from r_{replay}, r_{IKE}, and r_{HAP}; GAN showed its largest KRI contribution in spectral inflation $(r_{\mathrm{IKE}} = 0.27, r_{\mathrm{SPEC}} = 0.35)$. By contrast, KRI submetrics for RBM were near zero, including a minimal haplotype-collision signal $(r_{\mathrm{HAP}} = 0.05)$. For TLI (Fig. 1c, Supplementary Table 3), all three models showed low membership-inference performance $(r_{\mathrm{mia}} < 0.04)$. However, the rarity-collision component differs: RBM showed the highest $(r_{\mathrm{uniq}} = 0.79)$, while GAN was moderate $(r_{\mathrm{uniq}} = 0.27)$, and Genomator was noncontributing across H $(r_{\mathrm{uniq}} = 0)$.

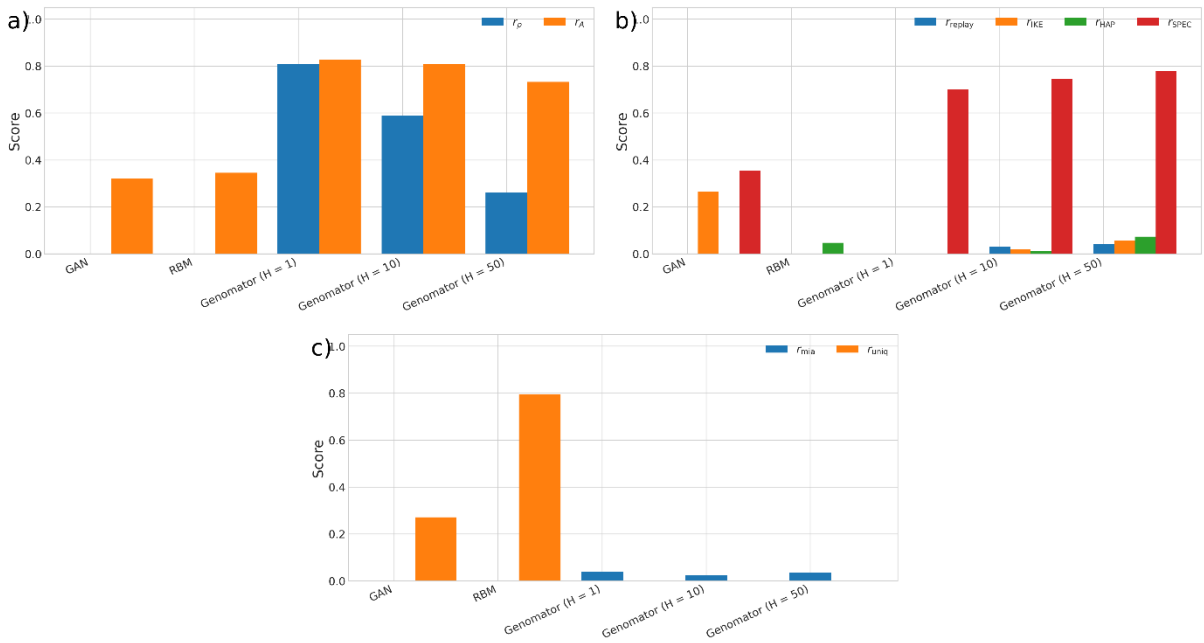

Figure 1. PRISM-G submetrics for the 10,000 SNPs panel dataset. **a)** Proximity leakage components: proximity quantile and adversarial distance signals. **b)** Kinship replay components: replay, internal kinship excess, micro haplotype collision, and spectral anomaly. **c)** Trait leakage components: membership-inference performance, uniqueness and rarity match.

On the 65,535 SNPs panel from chromosome 1, PLI submetrics for GAN and RBM indicated no leakage, whereas Genomator showed a monotonic decrease with H (Fig. 2a, Supplementary Table 4). For KRI (Fig. 2b, Supplementary Table 5), GAN displayed moderate spectral inflation ($r_{\mathrm{SPEC}} = 0.49$), while RBM showed a strong replay signal ($r_{\mathrm{replay}} = 0.74$). For Genomator, KRI submetrics varied with H: replay contributed at $H = 1$ ($r_{\mathrm{replay}} = 0.53$) and $H = 10$ ($r_{\mathrm{replay}} = 0.65$), whereas haplotype-collision and spectral inflation dominated at $H = 50$ ($r_{\mathrm{HAP}} = 0.34, r_{\mathrm{SPEC}} = 0.51$). For TLI (Fig. 2c, Supplementary Table 6), GAN and Genomator showed low membership-inference performance ($r_{\mathrm{mia}} < 0.08$) and negligible rarity-collision signals ($r_{\mathrm{uniq}} = 0$), whereas RBM exhibited an elevated rarity-collision component $\left(r_{\mathrm{uniq}} = 0.56\right)$.

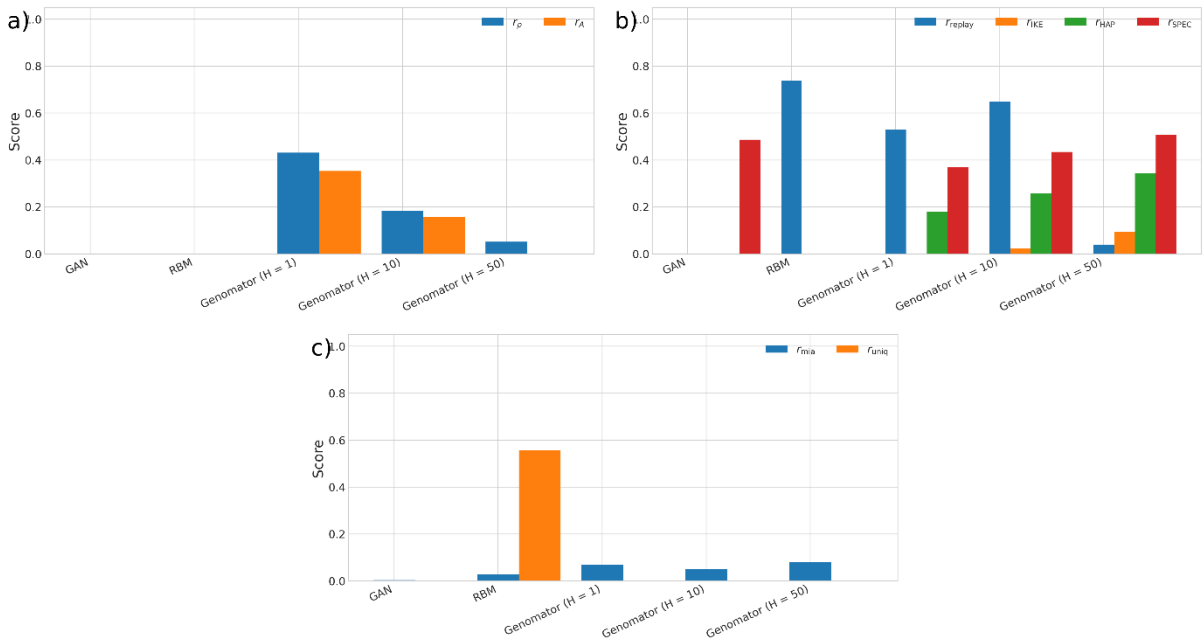

Figure 2. PRISM-G submetrics for the 65,535 SNPs panel dataset. **a)** Proximity leakage components: proximity quantile and adversarial distance signals. **b)** Kinship replay components: replay, internal kinship excess, micro haplotype collision, and spectral anomaly. **c)** Trait leakage components: membership-inference performance, uniqueness and rarity match.

Next, we computed the component-level PRISM-G scores. Figure 3 shows radar plots of the component scores for each generator. For the 10,000 SNPs panel, GAN and RBM displayed moderate proximity leakage $(0.32 < \mathrm{PLI} < 0.35)$, with kinship replay moderate for GAN $(\mathrm{KRI} = 0.35)$ and low for RBM $(\mathrm{KRI} = 0.05)$. Trait leakage was low for GAN $(\mathrm{TLI} = 0.27)$ but high for RBM $(\mathrm{TLI} = 0.79)$. Genomator presented the highest PLI at $H = 1$ $(\mathrm{PLI} = 0.83)$, with proximity leakage decreasing at $H = 10$ $(\mathrm{PLI} = 0.81)$ and $H = 50$ $(\mathrm{PLI} = 0.73)$, while kinship replay remained high $(0.70 < \mathrm{KRI} < 0.78)$ and trait leakage low $(0.02 < \mathrm{TLI} < 0.04)$. On the 65,535 SNPs panel, GAN and RBM showed negligible proximity leakage $(\mathrm{PLI} = 0)$. RBM displayed high kinship replay $(\mathrm{KRI} = 0.74)$, whereas GAN's replay reflected only mild spectral concentration $(\mathrm{KRI} = 0.49)$; GAN showed a negligible trait leakage signal $(\mathrm{TLI} = 0.003)$, while RBM was the highest $(\mathrm{TLI} = 0.56)$. Genomator's component profile compressed toward lower PLI as H increased $(0.05 < \mathrm{PLI} < 0.43)$, with mid-range kinship replay

$(0.51 < \mathrm{KRI} < 0.65)$ and low trait leakage $(0.05 < \mathrm{TLI} < 0.08)$ contributions. Overall, the component view reveals distinct failure modes: coordinate over-closeness (PLI) under tight constraints, relatedness inflation (KRI) for energy- and logic-based models, and trait-linked exposure (TLI) primarily through rare-variant collisions. Full component estimates are provided in Supplementary Tables 7-10.

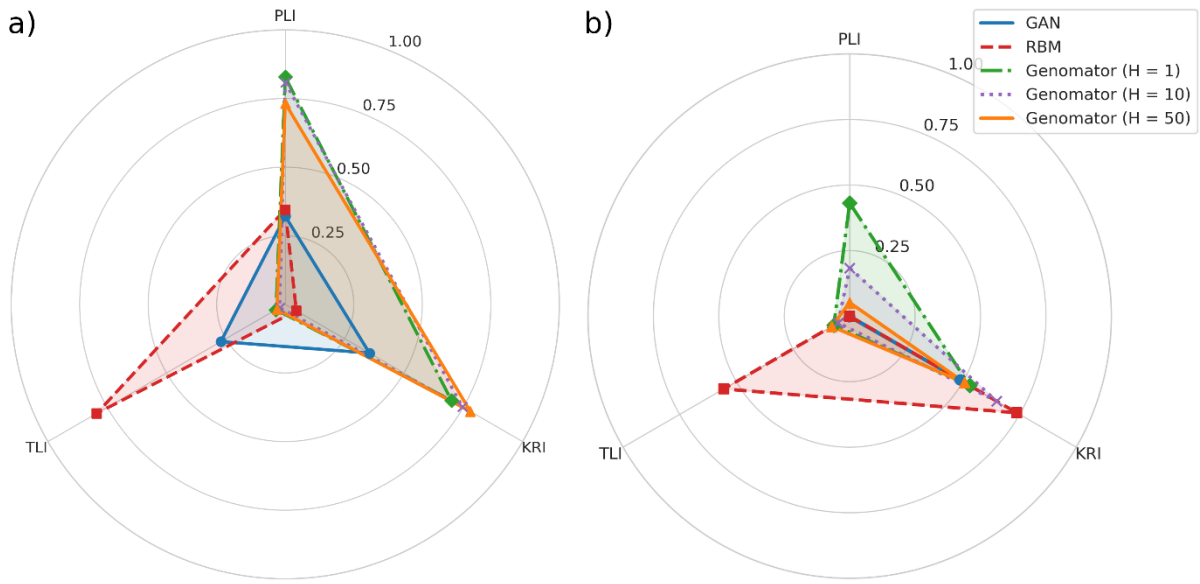

Figure 3. PRISM-G components. **a)** Metrics for the 10,000 SNPs panel dataset. **b)** Metrics for the 65,535 SNPs panel dataset.

PRIM-G scores and privacy risk assessment

Finally, we computed PRISM-G scores using safe and leaky datasets as anchors to place each generator on a common risk scale. Figure 4 and Supplementary Tables 7-10 report the raw and calibrated PRISM-G scores for all generators on the two SNP panels. On the 10,000 SNPs dataset from chromosome 15 (Fig. 4a), the safe binomial sampler anchored the lower bound at 0 $(\alpha = 0.04)$ and the leaky copycat at 100 $(\beta = 0.57)$. Among the learned models, GAN scored 46.13 $(R_{\mathrm{raw}} = 0.28)$ and RBM 60.46 $(R_{\mathrm{raw}} = 0.36)$, indicating moderate overall risk. Genomator yielded the highest scores on this panel, clustering between 77.06 and 78.64 across H $(0.447 < R_{\mathrm{raw}} < 0.455)$, consistent with the elevated

proximity and kinship leakage signals observed in the submetrics. Based on these results, the GAN-generated dataset is classified as **green** (safest), whereas the RBM- and Genomator-generated datasets are classified as **amber** (leaky) relative to the safe and leaky anchors.

On the 65,535 SNPs dataset from chromosome 1, the ordering shifted. GAN scored 27.17 ($R_{\text{raw}} = 0.16$), Genomator decreased with looser H constraints, scoring between 34.41 and 52.93 ($0.20 < R_{\text{raw}} < 0.31$), and RBM rose to 65.83 ($R_{\text{raw}} = 0.39$), the highest among the models. The safe ($\alpha = 0.006$) and leaky ($\beta = 0.58$) baselines again anchored the scale at 0 and 100, respectively. Compared with the chromosome 15 panel, the denser chromosome 1 SNP panel separated models more clearly: RBM's higher score is driven by kinship replay and rare-collision signals, whereas Genomator's score tracked its distance parameter, and GAN remained comparatively lower risk. These results label the GAN dataset **green** (safest), the Genomator dataset **green to amber** depending on the tuning parameter, and the RBM dataset **amber** (leaky), relative to the calibration anchors.

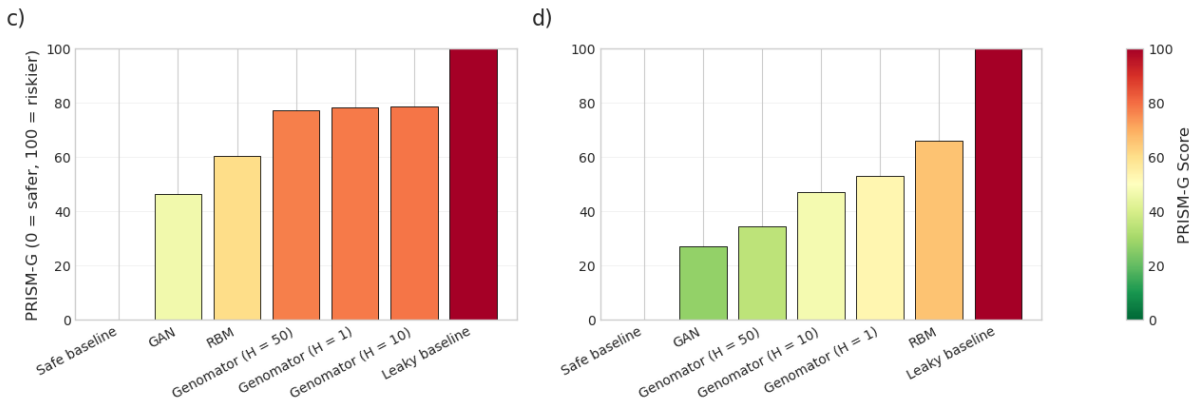

Figure 4. PRISM-G scores for synthetic data generators. **a)** 10,000 SNPs panel datasets. **b)** 65,535 panel datasets.

Discussion

In this study, we introduced PRISM-G, an integrated, model-agnostic framework for assessing privacy risks in synthetic genomic data across three complementary representations: genetic coordinate space (PLI), family structure (KRI), and trait-linked features derived from rare-variant content (TLI). By design, PRISM-G reveals distinct failure modes that single-view privacy metrics may miss.

Current evaluations of synthetic genomic data often lean on similarity-based diagnostics to argue that generated genomes are "far enough" from real sequences. However, such metrics have been shown to be insufficient as privacy guarantees, failing to address singling-out, linkability, and motivated-intruder scenarios (Ganev, 2024; Kaabachi et al., 2025). Yet distance alone does not establish resistance to concrete attacks or capture risk arising from linked information (Zhang et al., 2022): long-range familiar search in genealogy resources can re-identify individuals (Edge & Coop, 2020) and functional-genomics releases can leak attributes through cryptic channels and data linkage (Gürsoy et al., 2021; Harmanci & Gerstein, 2018). PRISM-G explicitly brings these factors into scope. Our proximity component (PLI) offers a calibrated view of coordinate closeness; the kinship component (KRI) shows replayed relatedness and spectral concentration that can facilitate long-rage familial search; and the trait component (TLI) targets exposure through rare-variant collisions and simple membership-inference advantages, both recognized vectors for identification or singling-out (Shringarpure & Bustamante, 2015). In doing so, PRISM-G aligns with and operationalizes emerging recommendations for domain-aware, task-linked evaluation and standardized reporting frameworks for synthetic medical data, moving beyond generic similarity indices (Pilgram et al., 2025; Zamzmi et al., 2025).

Our experiments on two marker panels (10,000 SNPs from chromosome 15 and 65,535 SNPs from chromosome 1) using the 1000 Genome Project as training dataset, showed that different generators concentrate risk along different axes as highlighted with prior attack-based evaluation of synthetic genomic data. Oprisnau et al. found that no single generator achieves both uniformly high utility and strong privacy and that some individuals remain vulnerable to membership inference even when

population-level distributions are well matched (Oprisanu et al., 2022). Using the GAN and RBM baselines from Yelmen et al. and a logic-based generator (Genomator) (Burgess et al., 2024.; Yelmen et al., 2023), PRISM-G component scores and submetrics reveal distinct patterns: at lower SNP density the GAN model exhibits a more balanced leakage profile, whereas at higher density it spreads samples sufficiently to avoid proximity and rare-variant collisions while still learning aspects of family structure. GANs are known to face mode-collapse which can produce synthetic data with limited diversity and, hence, hampering their effectiveness (Goodfellow, 2017). However, methods and settings that improve coverage can mitigate collapse and increase sample diversity (Kuo et al., 2023; Yelmen et al., 2023). Our observations align with improved coverage rather than collapse.

By contrast, our results suggest that RBM model tends to memorize rare variant patterns and family structures more strongly, elevating trait and kinship leakage signals in the PRISM-G score. Unintended memorization is a documented risk in generative models and can disproportionately affect uncommon samples, suggesting a plausible mechanism for our empirical findings (Carlini et al., 2019; Y. Chen et al., 2025). Aligned with our results, Genomator closely preserves global structure under tight distance constraints, elevating proximity and spectral-relatedness components while keeping rare variant content comparatively private; relaxing the constraint reduces proximity leakage but can intensify spectral relatedness, so overall risk is panel-dependent, consistent with the developers' observation that parameter choices modulate privacy risk.

Taken together, our results reinforce that genomic "privacy" is inherently multi-faceted rather than a single-number attribute. They also describe that: (i) membership-inference vulnerability and analytical utility do not move in lockstep across generators and datasets, and (ii) models that closely match population distributions can still expose specific individuals to inference risk (Adams et al., 2024).

Practically, three major key points emerge from our evaluation using PRISM-G. First, component scores are essential: the same integrated score can arise from different mechanisms, such as proximity over-closeness, inflated relatedness, or trait-linked uniqueness. Reporting PLI, KRI, and TLI metrics alongside the aggregate placement can guide mitigation (e.g., tuning distance constraints to reduce PLI; decorrelating or enforcing pedigree pruning to reduce KRI; down-weighting rare-variant patterns to reduce PLI). Second, calibration against simple anchors yields an interpretable 0-100 scale without assuming a specific adversary model; in practice we found these anchors bracketed the empirical range and made across-model comparisons straightforward. Third, our results reinforce that no single generator is uniformly safest: risk shifts with marker density, cohort composition, and model hyperparameters.

There are several limitations to note. PRISM-G does not provide a formal privacy guarantee (e.g., differential privacy); it is an empirical auditing framework whose conclusions depend on chosen features, thresholds, and baselines. Our coordinate view used PCA on common variants; alternative embeddings (e.g., UMAP or autoencoder latents) could change PLI sensitivity. KRI relies on GRM summaries and spectral patterns; very long-range haplotype structure or phasing artifacts may modulate these signals. TLI focuses on rare-variant collisions and a minimalist MIA; richer trait features (e.g., gene- or pathway-level burdens) or stronger attacks could alter rankings. Finally, our experiments were limited to subsets of 1KGP and two SNP panels; applying PRISM-G to broader, more diverse population cohorts will better stress-test data generators and clarify how the metrics behave across ancestries, marker densities, and cohort designs.

Building on our findings, PRISM-G suggests several directions. Methodologically, refining the algorithms and adding regularization to yield robust scores for genome-wide synthetic panels, and extending TLI to task-specific phenotypes and functional genomics signals would better align privacy risk with real analytic use cases. Empirically, large-scale tests on whole-genome data, with explicit

admixture and rare-variant stratification, would stress the corner cases where replay and uniqueness are most likely. On the mitigation side, combining PRISM-G with privacy-preserving training or post-generation filters (e.g., data augmentation, allele-frequency-aware filtering, pedigree de-duplication, or differential privacy mechanisms) could translate component insights into concrete safeguards.

Conclusions

Overall, PRISM-G offers a compact, calibrated ranking of privacy risk for synthetic genomic datasets while retaining interpretability across three complementary dimensions: PLI (proximity in genetic coordinate space), KRI (family structure and relatedness), and TLI (trait-linked features, particularly rare-variant exposure). Together, these components characterize where vulnerability arises, whether from over-closeness to real individuals, replayed kinship, or trait-driven uniqueness. Reporting both the integrated PRISM-G score and its component profiles is therefore informative: the single number supports deployment decisions and the choice of safeguards, while the submetrics explain why a model falls where it does on the 0–100 scale and point directly to targeted mitigations.

Funding

This work was supported by the Research Council KU Leuven project SHARE: Synthetic Human genomes for Advanced Research and Ethical data sharing (PDMT1/25/012).

Author contributions

A.C.R. conceptualized the study, developed the methodology and conducted the experiments. G.E. and Y.M. supervised this work. The original draft was written by A.C.R. A.C.R., G.E., Y.M. reviewed and edited the draft.

Declaration of interest

The authors declare that they have no competing interests

Code availability

Analysis scripts and codes are available on GitHub at https://github.com/alejocrojo09/prismg

References

Adams, T., Birkenbihl, C., Otte, K., Ng, H. G., Rieling, J. A., Näher, A.-F., Sax, U., Prasser, F., Fröhlich, H., & Alzheimer's Disease Neuroimaging Initiative. (2024). *On the Fidelity versus Privacy and Utility Trade-Off of Synthetic Patient Data*. Health Informatics. https://doi.org/10.1101/2024.12.06.24317239

Arora, A., Wagner, S. K., Carpenter, R., Jena, R., & Keane, P. A. (2025). The urgent need to accelerate synthetic data privacy frameworks for medical research. *The Lancet Digital Health*, *7*(2), e157–e160. https://doi.org/10.1016/S2589-7500(24)00196-1

Bonomi, L., Huang, Y., & Ohno-Machado, L. (2020). Privacy challenges and research opportunities for genomic data sharing. *Nature Genetics*, *52*(7), 646–654. https://doi.org/10.1038/s41588-020-0651-0

Burgess, M. A., Hosking, B., Reguant, R., Kaphle, A., O'Brien, M. J., Sng, M. F., Jain, Y., & Bauer, D. C. (2024). *Privacy-hardened and hallucination-resistant synthetic data generation with logic-solvers*.

Burnard, C., Mancheron, A., & Ritchie, W. (2025). Generating realistic artificial human genomes using adversarial autoencoders. *NAR Genomics and Bioinformatics*, *7*(3), lqaf101. https://doi.org/10.1093/nargab/lqaf101

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., & Song, D. (2019). *The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks* (No. arXiv:1802.08232). arXiv. https://doi.org/10.48550/arXiv.1802.08232

Chen, Y., Wang, S., Zou, D., & Ma, X. (2025). *SIDE: Surrogate Conditional Data Extraction from Diffusion Models* (No. arXiv:2410.02467). arXiv. https://doi.org/10.48550/arXiv.2410.02467

Chen, Y.-M., Hsiao, T.-H., Lin, C.-H., & Fann, Y. C. (2025). Unlocking precision medicine: Clinical applications of integrating health records, genetics, and immunology through artificial intelligence. *Journal of Biomedical Science*, *32*(1), 16. https://doi.org/10.1186/s12929-024-01110-w

Edge, M. D., & Coop, G. (2020). Attacks on genetic privacy via uploads to genealogical databases. *eLife*, *9*, e51810. https://doi.org/10.7554/eLife.51810

Ganev, G. (2024). *Synthetic Data, Similarity-based Privacy Metrics, and Regulatory (Non-)Compliance* (No. arXiv:2407.16929). arXiv. https://doi.org/10.48550/arXiv.2407.16929

Goodfellow, I. (2017). *NIPS 2016 Tutorial: Generative Adversarial Networks* (No. arXiv:1701.00160). arXiv. https://doi.org/10.48550/arXiv.1701.00160

Goudet, J., Kay, T., & Weir, B. S. (2018). How to estimate kinship. *Molecular Ecology*, *27*(20), 4121–4135. https://doi.org/10.1111/mec.14833

Gürsoy, G., Brannon, C. M., Navarro, F. C. P., & Gerstein, M. (2021). FANCY: Fast estimation of privacy risk in functional genomics data. *Bioinformatics*, *36*(21), 5145–5150. https://doi.org/10.1093/bioinformatics/btaa661

Gürsoy, G., Li, T., Liu, S., Ni, E., Brannon, C. M., & Gerstein, M. B. (2022). Functional genomics data: Privacy risk assessment and technological mitigation. *Nature Reviews Genetics*, *23*(4), 245–258. https://doi.org/10.1038/s41576-021-00428-7

Harmanci, A., & Gerstein, M. (2018). Analysis of sensitive information leakage in functional genomics signal profiles through genomic deletions. *Nature Communications*, *9*(1), 2453. https://doi.org/10.1038/s41467-018-04875-5

Kaabachi, B., Despraz, J., Meurers, T., Otte, K., Halilovic, M., Kulynych, B., Prasser, F., & Raisaro, J. L. (2025). A scoping review of privacy and utility metrics in medical synthetic data. *Npj Digital Medicine*, *8*(1), 60. https://doi.org/10.1038/s41746-024-01359-3

Kenneweg, P., Dandinasivara, R., Luo, X., Hammer, B., & Schönhuth, A. (2025). Generating synthetic genotypes using diffusion models. *Bioinformatics*, *41*(Supplement_1), i484–i492. https://doi.org/10.1093/bioinformatics/btaf209

Kuo, N. I.-H., Garcia, F., Sönnerborg, A., Böhm, M., Kaiser, R., Zazzi, M., Polizzotto, M., Jorm, L., & Barbieri, S. (2023). Generating synthetic clinical data that capture class imbalanced distributions with generative adversarial networks: Example using antiretroviral therapy for HIV. *Journal of Biomedical Informatics*, *144*, 104436. https://doi.org/10.1016/j.jbi.2023.104436

Madsen, B. E., & Browning, S. R. (2009). A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. *PLoS Genetics*, *5*(2), e1000384. https://doi.org/10.1371/journal.pgen.1000384

Mani, S., Lalani, S. R., & Pammi, M. (2025). Genomics and multiomics in the age of precision medicine. *Pediatric Research*, *97*(4), 1399–1410. https://doi.org/10.1038/s41390-025-04021-0

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, *26*(22), 2867–2873. https://doi.org/10.1093/bioinformatics/btq559

Mendes, J. M., Barbar, A., & Refaie, M. (2025). Synthetic data generation: A privacy-preserving approach to accelerate rare disease research. *Frontiers in Digital Health*, *7*, 1563991. https://doi.org/10.3389/fdgth.2025.1563991

Oldoni, F., Kidd, K. K., & Podini, D. (2019). Microhaplotypes in forensic genetics. *Forensic Science International: Genetics*, *38*, 54–69. https://doi.org/10.1016/j.fsigen.2018.09.009

Oprisanu, B., Ganev, G., & Cristofaro, E. D. (2022). *On Utility and Privacy in Synthetic Genomic Data* (No. arXiv:2102.03314). arXiv. https://doi.org/10.48550/arXiv.2102.03314

Patterson, N., Price, A. L., & Reich, D. (2006). Population Structure and Eigenanalysis. *PLoS Genetics*, *2*(12), e190. https://doi.org/10.1371/journal.pgen.0020190

Pilgram, L., Dankar, F. K., Drechsler, J., Elliot, M., Domingo-Ferrer, J., Francis, P., Kantarcioglu, M., Kong, L., Malin, B., Muralidhar, K., Myles, P., Prasser, F., Raisaro, J. L., Yan, C., & El Emam, K. (2025). A consensus privacy metrics framework for synthetic data. *Patterns*, 101320. https://doi.org/10.1016/j.patter.2025.101320

Seidman, D. N., Shenoy, S. A., Kim, M., Babu, R., Woods, I. G., Dyer, T. D., Lehman, D. M., Curran, J. E., Duggirala, R., Blangero, J., & Williams, A. L. (2020). Rapid, Phase-free Detection of Long Identity-by-Descent Segments Enables Effective Relationship Classification. *The American Journal of Human Genetics*, *106*(4), 453–466. https://doi.org/10.1016/j.ajhg.2020.02.012

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). *Membership Inference Attacks against Machine Learning Models* (No. arXiv:1610.05820). arXiv. https://doi.org/10.48550/arXiv.1610.05820

Shringarpure, S. S., & Bustamante, C. D. (2015). Privacy Risks from Genomic Data-Sharing Beacons. *The American Journal of Human Genetics*, *97*(5), 631–646. https://doi.org/10.1016/j.ajhg.2015.09.010

Stark, Z., Glazer, D., Hofmann, O., Rendon, A., Marshall, C. R., Ginsburg, G. S., Lunt, C., Allen, N., Effingham, M., Hastings Ward, J., Hill, S. L., Ali, R., Goodhand, P., Page, A., Rehm, H. L., North, K. N., & Scott, R. H. (2025). A call to action to scale up research and clinical genomic data sharing. *Nature Reviews Genetics*, *26*(2), 141–147. https://doi.org/10.1038/s41576-024-00776-0

Venturini, P., Faria, P. L., & Cordeiro, J. V. (2025). AI and omics technologies in biobanking: Applications and challenges for public health. *Public Health*, *243*, 105726. https://doi.org/10.1016/j.puhe.2025.105726

Wharrie, S., Yang, Z., Raj, V., Monti, R., Gupta, R., Wang, Y., Martin, A., O'Connor, L. J., Kaski, S., Marttinen, P., Palamara, P. F., Lippert, C., & Ganna, A. (2023). HAPNEST: Efficient, large-scale generation and evaluation of synthetic datasets for genotypes and phenotypes. *Bioinformatics*, *39*(9), btad535. https://doi.org/10.1093/bioinformatics/btad535

Yale, A., Dash, S., Dutta, R., Guyon, I., Pavao, A., & Bennett, K. P. (2019). *Privacy Preserving Synthetic Health Data*.

Yelmen, B., Decelle, A., Boulos, L. L., Szatkownik, A., Furtlehner, C., Charpiat, G., & Jay, F. (2023). Deep convolutional and conditional neural networks for large-scale genomic data generation. *PLOS Computational Biology*, *19*(10), e1011584. https://doi.org/10.1371/journal.pcbi.1011584

Yelmen, B., Decelle, A., Ongaro, L., Marnetto, D., Tallec, C., Montinaro, F., Furtlehner, C., Pagani, L., & Jay, F. (2021). Creating artificial human genomes using generative neural networks. *PLOS Genetics*, *17*(2), e1009303. https://doi.org/10.1371/journal.pgen.1009303

Zamzmi, G., Subbaswamy, A., Sizikova, E., Margerrison, E., Delfino, J. G., & Badano, A. (2025). Scorecard for synthetic medical data evaluation. *Communications Engineering*, *4*(1), 130. https://doi.org/10.1038/s44172-025-00450-1

Zhang, Z., Yan, C., & Malin, B. A. (2022). Membership inference attacks against synthetic health data. *Journal of Biomedical Informatics*, *125*, 103977. https://doi.org/10.1016/j.jbi.2021.103977