PRISM-G: an interpretable privacy scoring framework 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

*Corresponding author. Email: alejandro.correarojo@esat.kuleuven.be[1], gokhan.ertaylan@vito.be[2]

Abstract

Synthetic genomic data promises broader access, but unresolved privacy risks persist. In Europe, these risks increasingly hinder cross-border use of national genomic resources due to limitations in trust and legal interoperability rather than scientific demand. At the same time, privacy risk is not uniformly distributed: leakage driven by relatedness structure and rare-variant uniqueness can disproportionately affect underrepresented or vulnerable populations when shared data are misused or linked with external resources, making transparent, domain-aware measurement of privacy exposure central to responsible governance. We introduce PRISM-G, a model-agnostic framework that quantifies privacy exposure in synthetic genome data cross three complementary components: proximity to real genomes in genetic-coordinate space, replay of familial or population-structure patterns, and trait-linked exposure via rare variants and membership-inference signals. These components are normalized and combined through a risk-averse aggregation into a single 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, demonstrating that a single similarity-based metric is sufficient to characterize genomic privacy risk.

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 [1]. 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 [2,3]. However, the fact that personal genetic and health-related data is sensitive (Conduah et al., 2025; Evans & Jarvik, 2018; Shabani & Marelli, 2019), coupled to the fact that genomes are inherently re-identifiable [7], make broad data sharing a persistent technological, ethical, and legal challenge because demonstrated privacy risks and stringent regulatory requirements [8,9]. This challenge is particularly acute in Europe, where the promise of cross-border genomic research and clinical translation is tightly coupled to governance, legal interoperability, and public trust. Multiple European initiatives, including the European Health Data Space (EHDS), aim to enable secure secondary use of health and genomic data at scale, yet operational data sharing has repeatedly stalled due to disagreements over acceptable risk, divergent national interpretations of data-protection obligations, and the lack of widely accepted technical evidence that released data cannot be linked back to individuals [10–12].

As a response to these challenges, synthetic data generation has emerged as a potential strategy to expand data availability while reducing privacy risks and preserving analytical utility without direct reliance on real personal records [13]. In genetic research, artificial genome datasets are considered a promising tool, as many analyses depend on large-scale individual genetic variation that synthetic data could help provide for population genetics, rare disease studies, individual disease risk estimation, and

improved representation of underrepresented populations [14–16]. Moreover, synthetic data may facilitate genomic data sharing across collaborators and developers by easing regulatory constraints, while remaining subject to ongoing ethical and legal considerations [17,18].

A growing set of methods exists for generating synthetic genomic data. Traditional haplotype resampling combined with coalescent simulations (e.g., HAPNEST) provide scalable, reference-guided genotype synthesis and quality evaluation [16]. More recently, deep generative models, such as generative adversarial networks (GANs), variational autoencoders (VAEs), restricted Boltzmann machines (RBMs), and diffusion models, aim to capture linkage disequilibrium, population structure and higher-order genomic dependencies [19,20]. In parallel, logic-based generators (e.g., Genomator) formulate data synthesis as a constrained optimization problem to enforce biological plausibility and privacy constraints [21]. Although these approaches show promise in producing faithful synthetic cohorts, generative modeling for genomics remains an active and maturing area of research, and synthetic genomic data is only beginning to emerge as a practical substrate for population genetics, method development, and downstream modeling (Battey et al., 2021; Wang et al., 2025).

Despite the demonstrated feasibility of synthetic genomic data generation, systematic evaluation of privacy risk remains a central challenge. Current practices rely mainly on similarity-based metrics, such as nearest-neighbor or Hamming-distance proximity between synthetic and real genomes, as proxies for identifiability [16,21,24,25], often complemented by Membership Inference Attacks (MIAs) that test whether individuals can be distinguished between training and holdout data [26]. However, these approaches lack a unified interpretive framework and have sometimes led to the assumption that synthetic genomic data are “privacy-preserving by default”, despite evidence that disclosure risk varies substantially across methods [27]. Moreover, models that accurately reproduce population-level distributions may still expose individuals to re-identification or membership detection through data linkage, genealogical inference, or trait prediction using external reference data and functional

genomics [8,28,29]. Consequently, the absence of standardized, domain-aware evaluation protocols limits meaningful comparison between generators and complicates risk-based data sharing [30], and regulatory assessment under legal frameworks such as the General Data Protection Regulation (GDPR) [31].

Genomic data are intrinsically multi-representational. The same individual can be described in coordinate space capturing population structure (e.g., eigenanalysis of common variants) [32], in kinship space reflecting familial relatedness through genetic relationship matrices (GRMs), and identical-by-descent (IBD) measures [33], and in trait-linked feature space summarizing biologically meaningful signals such as rare-variant burdens [34]. Each of these representations encodes distinct aspects of personal information and reflects different ways in which individuals relate to populations, families, and traits.

For this reason, privacy risk in synthetic genomic data cannot be assessed solely by how closely individual synthetic genomes resemble real ones. Even when individual-level similarity is low, privacy leakage may arise if synthetic cohorts preserve relational structure shaped by shared ancestry or reproduce feature-level signals linked to distinctive genetic characteristics. Evaluating privacy risk therefore requires accounting for multiple, interacting pathways through which personal information may be exposed, beyond simple distance-based measures.

Building on these representations we present the **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: (1) **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; (2) **Kinship replay (KRI)** measures whether the synthetic cohort

inadvertently recreates family structure or long-range dependence (e.g., close cousins, inflated relatedness, repeated haplotype segments); and (3) **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 to ensure interpretability across datasets. A graphical description of the PRISM-G framework is illustrated in Figure 1.

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 [35].

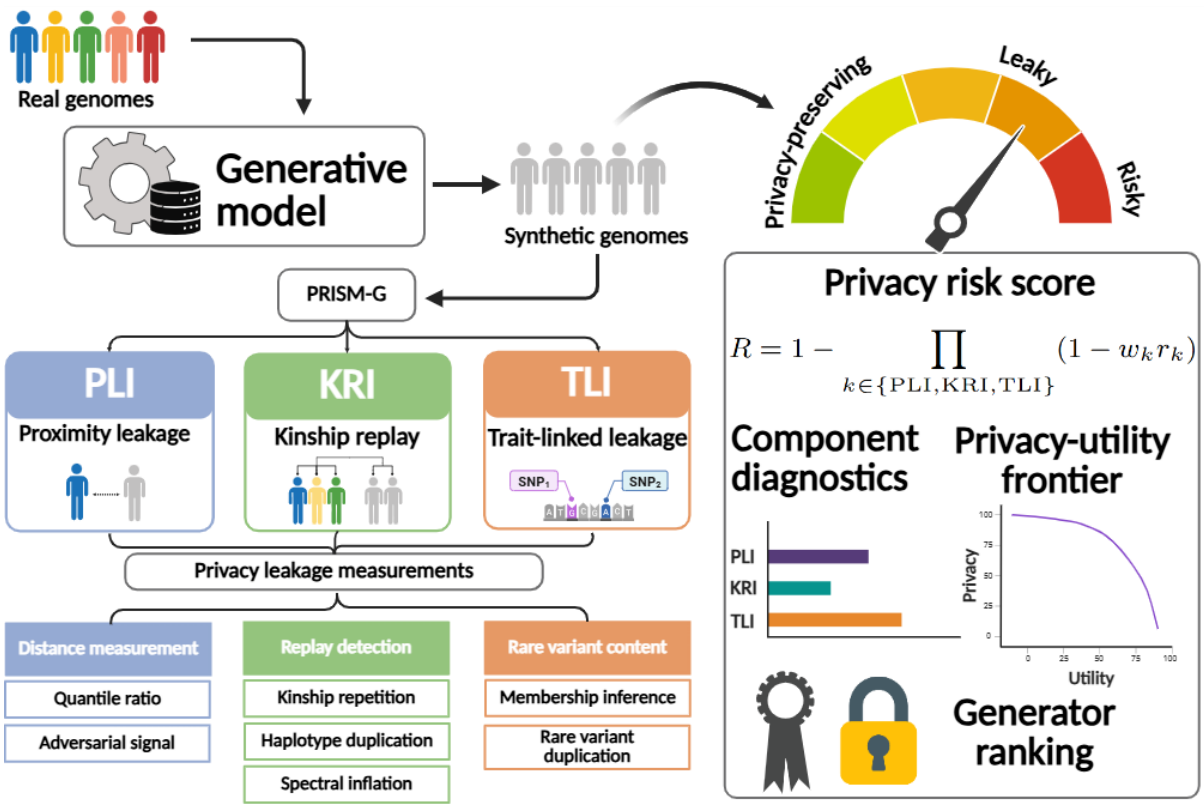

Figure 1. Overview of the PRISM-G privacy scoring framework.

Algorithmic basis and related work

As previously mentioned, early evaluations of synthetic genomic data often relied on similarity indices, such as distance to the nearest real neighbor, as proxy for identity risk. Yet genomic privacy risks are not limited to pointwise closeness: (1) population structure and relatedness create identifiable patterns that enable familial search and relationship inference, and (2) functional profiles and rare variants can reveal membership or traits even when overall distances look innocuous. Foundational work on principal component analysis (PCA) provides compact genetic axes that reveal population structure or demographic variation, while kinship estimation methods, such as GRM, operationalize relatedness at scale [32,36,37]. 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 [27,38]. Taking this into account, PRISM-G's components are designed to operationalize these documented risks using established genomic methodology: coordinates for proximity, GRM and short-haplotype patterns for relatedness replay, and MIA signals plus rare-variant counts for trait-linked exposure. Together, these components are aimed at capturing complementary and documented pathways through which personal information can leak from synthetic genome data. 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
Proximity leakage index (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.)
Kinship replay index (KRI)	Captures relatedness using GRM, tests for excess kinship against a safe baseline, scans for	PRISM-G combines four baseline-corrected signals: relatedness surplus, kin-edge	[33,36,37,40–42]

	short haplotype sharing via short unphased genotype window, and flags spectral anomalies by detecting inflation of the GRM's leading eigenvalue.	replay, micro-haplotype collisions, and spectral concentration, into a novel aggregate that detects replayed familial structure and long-range dependence.	
Trait-linked leakage index (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-occurrence of uncommon variants. Together the results are mapped into a single score.	[26,34,38]

Table 1. Conceptual overview of the PRISM-G components, the genomic representations they probe, and the types of privacy leakage they are designed to detect. Formal definitions and implementation details are provided in the Materials and Methods section.

PRISM-G methodology

This section describes the formal definitions, estimation procedures, and calibration steps used to compute PRISM-G. We first define the three-component metrics, PLI, KRI, and TLI, each corresponding to a distinct genomic representation. We then describe how these components are aggregated and calibrated into a unified risk score, followed by robustness analyses for assessing sensitivity to calibration choices, and ranking risk for genomic data generators.

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. Additional details on the methods used for component estimation can be found in the Supplementary Material.

Proximity leakage index

The PLI metric asks whether any synthetic genome lies unusually close to a real individual in genetic coordinate space, beyond what would be expected from population structure alone. For computation, we project real and synthetic samples into a low-dimensional coordinate space using PCA of genetic variants. For each synthetic sample, we evaluate its distance to the nearest real individual and compare the resulting distribution to a reference baseline. Two complementary terms capture this signal: a lower-tail quantile ratio measuring over-closeness and an adversarial proximity score reflecting nearest-neighbor concentration. These terms are combined into a normalized PLI score as defined below.

Genotypes are embedded into a d dimensional space (principal components computed after standardizing genotypes using allele frequencies estimated from R_{tr}) [32]. 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 as

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

and the baseline nearest real distance to other holdout samples as

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

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 the 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 $\text{A} = \frac{1}{n_{\text{ho}}} \sum_i a_i$. We map this to a risk as $r_{\text{A}} = \max\left\{0, \frac{0.5 - \text{A}}{0.5}\right\}$. The PLI is calculated as

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

Kinship replay index

The KRI metric evaluates whether synthetic datasets reproduce patterns of relatedness present in the real cohort, such as familiar relationships or population-level correlation structure. Unlike PLI, which focuses on individual proximity, KRI captures relational privacy risk that may persist even when no

single synthetic genome closely matches a real individual. KRI aggregates several complementary diagnostics, each targeting a different manifestation of relatedness replay. These include replay of close kinship pairs, excess internal relatedness, haplotype-level collisions, and inflation of global correlation structure. We compute GRMs for R_{ho} and S after standardizing genotypes by training set allele frequencies as described by VanRaden, yielding K^{ho} and K^{S} [37]. KRI aggregates four signals, each baseline corrected against a safe or bootstrap reference.

First, the **replay** component (r_{replay}) evaluates whether the synthetic dataset K^{S} reproduces the distribution of close-relatives relationships observed in the real cohort K^{ho}. For a kinship threshold θ, we measure the similarity M between the real and synthetic close-kin spectra using the Jensen-Shannon (JS) divergence as measurement (Supplementary Material). 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 (Supplementary Material). Let $f(\theta) = \Pr[K_{ij}^{S} \geq \theta, i < j]$. Using a holdout bootstrap set to estimate $f_0(\theta)$, we 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[42]. 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 (Supplementary Material). After subtracting a matched baseline c_0, we set

$$r_{\mathrm{HAP}} = \frac{c - c_0}{c_{\max} - c_0} \tag{8}$$

with $c_{\max}$ taken as the maximum rate across all windows.

Fourth, the **spectral inflation** (r_{SPEC}) captures 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 (Supplementary Material), define

$$r_{\mathrm{SPEC}} = 1 - \frac{s_0}{s}. \tag{9}$$

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 privacy risk arising from genetic features associated with phenotypes or distinctive individual characteristics, with a particular focus on rare genetic variants. Such features may enable attribute inference or membership attacks even in the absence of close genetic similarity. We quantify TLI using two complementary signals: performance of a simple MIA and the presence of rare-variant collisions between synthetic and real individuals. These terms reflect different ways in which trait-linked information may be exposed and are combined into a normalized TLI score as described below. Let c_i denote trait features for individual i (rare variant component burdens for chromosome sets)[34].

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

$$r_{\mathrm{mia}} = \max\left\{0, \frac{0.5 - \mathrm{A}}{0.5}\right\}, \tag{12}$$

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 hypothesis driven by training allele frequencies[38]. Define the rare set $R = \{j : MAF_j^{\mathrm{train}} < \tau\}$. For 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 (Supplementary Material) and set

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

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. Because raw component scores are not directly comparable across datasets or generators, we aggregate them into a single cohort-level risk score. To do so, we use a risk-averse, "OR-like" aggregation that preserves monotonicity and ensures that a single high-risk component cannot be masked by lower values in the others.

Specifically, let $r_{\text{PLI}}, r_{\text{KRI}}, r_{\text{TLI}} \in [0, 1]$ denote the component scores and $w_{\text{PLI}}, w_{\text{KRI}}, w_{\text{TLI}} \geq 0$ weights that sum to one. The aggregated raw risk is defined as:

$$R_{\text{raw}} = 1 - \prod_{k \in \{\text{PLI, KRI, TLI}\}} (1 - w_k r_k). \tag{16}$$

This construction has three practical advantages. First, it is **monotone** in every component: increasing any r_k 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 term is omitted from KRI's internal maximum without renormalizing other components. This preserves comparability across datasets.

Calibration and estimation of weights

Because raw risk scores are not directly comparable across datasets or generators, PRISM-G uses simple reference generators to anchor the risk scale. This calibration step ensures that scores are interpretable and comparable without assuming a specific adversary model. We calibrate R_{raw} using two reference generators that anchor the lower and upper ends of the scale:

1. A **safe baseline** (A_{Safe}) that preserves allele frequencies but intentionally removes dependence structure (for example, an allele-frequency–matched marginal sampler).
2. One or more **leaky baselines** (A_{Leaky}) that intentionally overfit structure (for example, copycat, kinship-preserving variants).

To prevent the estimator from "spiking" on a single component and inflating overall risk, we calibrate the weights with a ridge-regularized fit to the anchor baselines. Concretely, we choose w to make the aggregated risk $R(\cdot; w)$ hit target values at the anchors while staying close to equal weights w_0:

$$\min_{w \in \Delta} (R(A_{\text{Safe}}; w) - \varepsilon)^2 + \left(R\left(A_{\text{Leaky}}; w\right) - \lambda\right)^2 + \gamma \|w - w_0\|_2^2 \tag{17}$$

where the loss function is subject to $w_m \geq 0$ and $\sum_m w_m = 1$.

Using Equation 17, we define three hyperparameters to calibrate the scale of R_{raw}. The safe target ε anchors $R(A_{\text{Safe}}; w)$ to a small band (0.01–0.1), which avoids forcing any component weight to vanish

purely to hit zero exactly. The leaky target λ (0.75-0.90) anchors $R\left(A_{\text{Leaky}}; w\right)$ high on the scale while avoiding saturation, leaving headroom for cases that are even leakier. Finally, the penalty strength γ $(10^{-4} - 10^{-2})$ shrinks w toward the prior w_0, balancing anchor fit with stability and discouraging spiky, single-component solutions.

Estimation of PRISM-G score

For each baseline we compute the distribution of R_{raw}. Let α denote the calibrated lower anchor of R_{raw} under the safe baseline, and let β denote the upper anchor, under the leaky baseline. We then map the observed raw risk for a model of interest to a standardized 0–100 scale:

$$\text{PRISM-G} = 100 \times \frac{R_{\text{raw}} - \alpha}{\beta - \alpha}. \tag{18}$$

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.

Finally, to assess the robustness of the PRISM-G scores, we introduce a ranking stability test using Kendall's rank correlation coefficient. Using bootstrapping, let $\pi^{(b)}$ be the ranking at bootstrap draw b, and choose a baseline ranking $\pi^{(0)}$. The stability test is defined as

$$\bar{\tau} = \frac{1}{B}\sum_{b=1}^{B} \tau(\pi^{(b)}, \pi^{(0)}), \tag{19}$$

where $\bar{\tau}$ is the mean Kendall rank coefficient, and B the total number of bootstraps. The permutation p-value is obtained by comparing $\bar{\tau}$ to $\tau(\pi^{(\mathrm{perm})}, \pi^{(0)})$.

Grid search evaluation

Finally, for hyperparameter tuning, we perform a grid search over the calibration parameters to maximize $\bar{\tau}$ for PRISM-G score estimation. Given grids $\varepsilon \in \mathrm{E}$, $\lambda \in \Lambda$, $\gamma \in \Gamma$, we evaluate each triple $(\varepsilon, \lambda, \Gamma)$ and select the setting that yields the highest $\bar{\tau}$:

1. Fit w^* by the objective function provided in Equation 17.
2. Compute PRISM-G for all candidate datasets/models.
3. Bootstrap anchors to obtain rank samples and compute Kendall's $\bar{\tau}$ stability.
4. Return $\varepsilon, \lambda, \Gamma$ that maximizes $\bar{\tau}$.

Estimation of the privacy-utility trade-off

The calibrated PRISM-G score provides a compact summary of privacy exposure, while component diagnostics reveal the mechanisms underlying potential leakage. Because generative models often trade fidelity and utility for privacy protection, PRISM-G can be paired with task-specific utility metrics to analyze privacy–utility trade-offs, for example, by generating a Pareto frontier. This enables systematic comparison between models and facilitates the identification of generators that achieve favorable balances between analytical utility and privacy protection.

Evaluation procedure

Generative models

We evaluated PRISM-G on synthetic cohorts generated from three model families: (1) **GANs**, (2) **RBMs** as introduced by Yelmen and colleagues, and (3) **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 layers 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[20]. 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., 2025).

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.

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. The estimation of component metrics was performed over 1,000 bootstraps replicas.

For PLI metrics, we computed genetic coordinates using 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) less than 0.001, with allele frequencies estimated from the training set. Membership inference risk was computed as described in the PRISM-G methodology section.

For the initial estimation of raw values, we calibrated weights using hyperparameters $\varepsilon = 0.02$, $\lambda = 0.75$, and $\gamma = 10^{-3}$, and performed ranking stability test over 1,000 bootstraps. Finally, for grid search, we evaluated $\varepsilon \in \{0.02, 0.04, 0.06, 0.08, 0.10\}$, $\lambda \in \{0.75, 0.80, 0.85, 0.90\}$, and $\gamma \in \{10^{-4}, 10^{-3}, 10^{-2}\}$ over 1,000 bootstraps.

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.

Utility evaluation and privacy-utility trade-offs

To evaluate the privacy-utility trade-off of the synthetic genomic datasets, we assessed their performance in a genetic ancestry inference task. Population structure was first estimated from the real training cohort using PCA. The first ten principal components were retained to define the genetic coordinate space used for downstream analyses. Synthetic genomes were then projected onto the same PCA embedding to ensure comparability between real and synthetic samples.

Utility was evaluated using the superpopulations labels provided by the 1KGP datasets, including African (AFR), European (EUR), East Asian (EAS), South Asian (SAS), and Admixed American (AMR) populations. A random forest classifier was trained on the real training samples using the ten principal components and the corresponding superpopulation labels and then applied to the synthetic datasets to predict ancestry assignments.

We quantified population-structure preservation using two complementary metrics. First, Procrustes similarity was computed between the centroids of real and synthetic ancestry clusters in the PCA space, providing a measure of geometric alignment between population structures. Second, we measured the JS similarity between the distributions of predicted ancestry labels in the synthetic data and the corresponding distributions in the real holdout data. JS similarity was calculated using the proportions of individuals assigned to each superpopulation group.

Results

Overview of PRISM-G components

We first evaluated submetric components for each synthetic dataset to illustrate how different generators exhibit distinct privacy leakage mechanisms prior to aggregation into PRISM-G scores. Figures 1 and 2, together with Supplementary Tables 1-4, summarize the results for the two marker panels. For the 10,000 SNP panel from Chromosome 15 (Fig. 1a; Supplementary Table 1), PLI submetrics indicated that GAN and RBM showed little proximity-based leakage, with quantile ratios above unity, yielding $r_\rho = 0$, and only modest adversarial proximity above chance ($r_{\mathrm{A}} < 0.35$). In contrast, Genomator exhibited a tunable proximity profile ($H \in \{1, 10, 50\}$). Under tight constrains synthetic samples lay unusually close to holdout individuals, while increasing the constraint progressively reduced proximity-based exposure.

Kinship replay index submetrics (Fig. 1b; Supplementary Table 2) revealed distinct patterns across generators. Genomator's kinship risk was driven primarily by spectral inflation ($0.75 < r_{\mathrm{SPEC}} < 0.81$), with smaller contributions from the replay component (r_{replay}), internal kinship excess (r_{IKE}), and haplotype-collision terms (r_{HAP}). GAN showed moderate inflation in r_{SPEC} and r_{IKE} whereas RBM exhibited minimal kinship-related signals across all submetrics.

For TLI (Fig. 1c; Supplementary Table 3), membership inference performance remained low for all models ($r_{\mathrm{mia}} < 0.04$). However, rare variant uniqueness differed substantially: RBM showed the strongest rare-variant collision signal ($r_{\mathrm{uniq}} = 0.79$), GAN exhibited moderate uniqueness ($r_{\mathrm{uniq}} = 0.27$), and Genomator did not contribute across constraint setting ($r_{\mathrm{uniq}} = 0.00$).

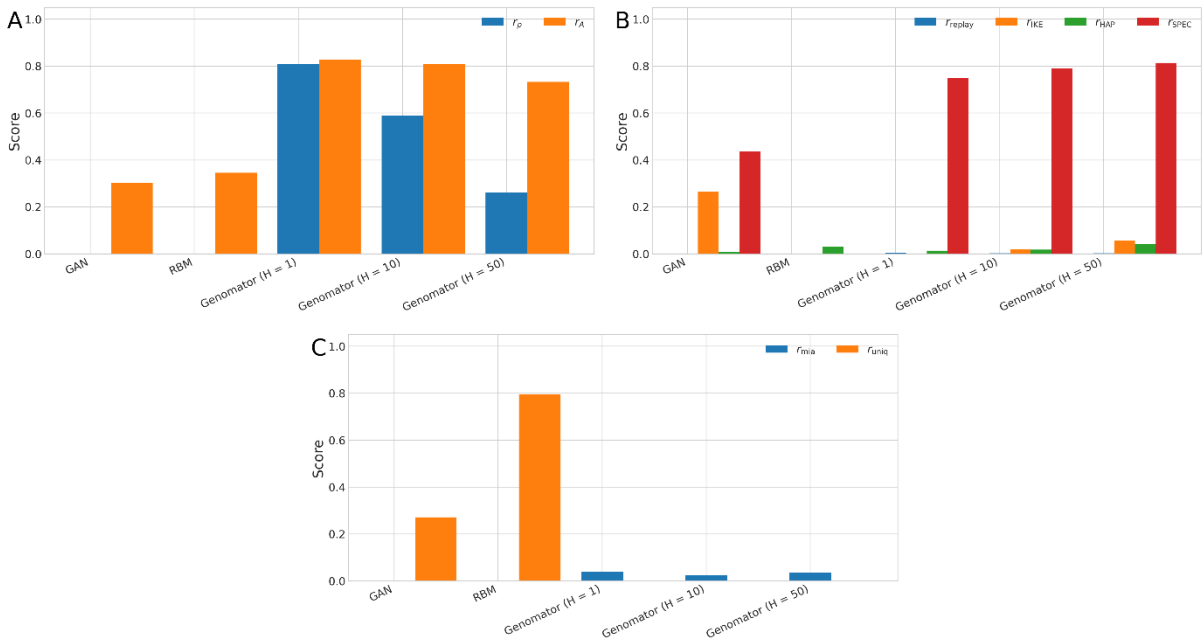

Figure 1. PRISM-G submetrics for the 10,000 SNPs panel dataset. Each barplot shows **A)** Proximity leakage components: proximity quantile and adversarial distance signals. **B)** Kinship replay components: replay, internal kinship excess, micro haplotype collision, and spectral anomaly. **C)** Trait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

On the 65,535 SNP panel from Chromosome 1, PLI submetrics indicated little proximity-based leakage for GAN and RBM, whereas Genomator showed a monotonic reduction in proximity exposure as the constraint parameter H increased (Fig. 2a; Supplementary Table 4). In contrast, KRI submetrics revealed distinct generator-specific patterns. GAN exhibited moderate spectral inflation ($r_{\mathrm{SPEC}} = 0.54$), RBM showed negligible kinship leakage across all KRI submetrics, and Genomator displayed a constraint-dependent shift, with haplotype-collision and spectral-relatedness signals (r_{HAP} and r_{SPEC}) dominating at higher H values.

For TLI (Fig. 2c; Supplementary Table 6), GAN and Genomator showed low membership-inference signals (r_{mia}) and negligible rare-variant uniqueness (r_{uniq}) signals, whereas RBM exhibited elevated rare-variant uniqueness component ($r_{\mathrm{uniq}} = 0.56$).

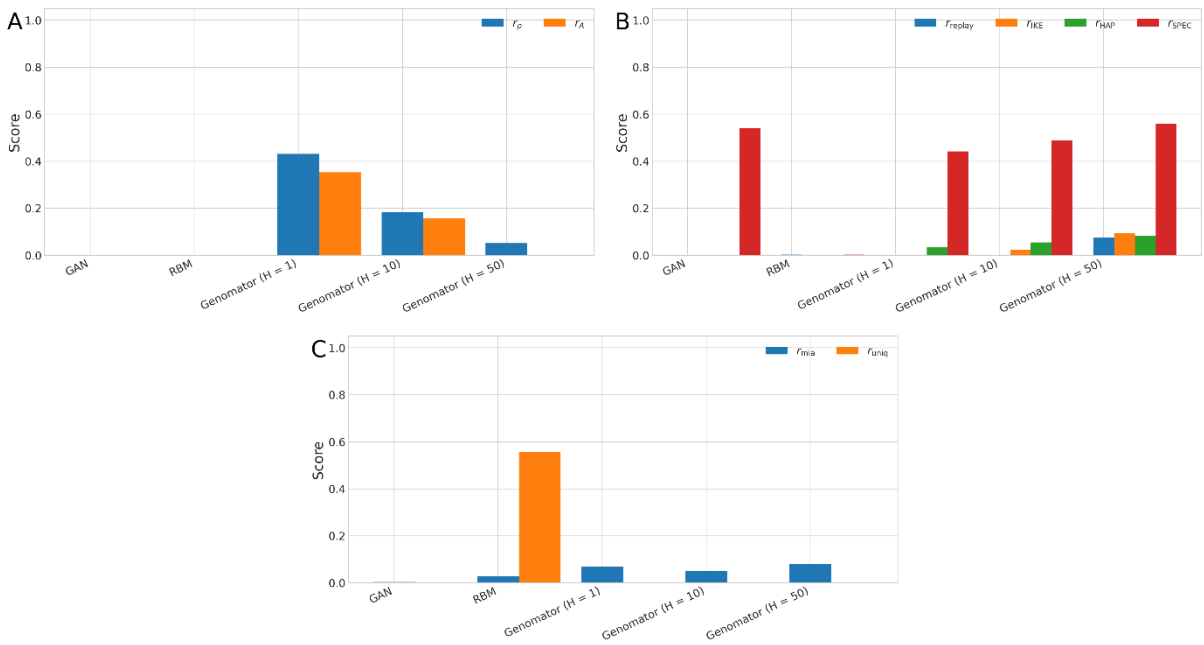

Figure 2. PRISM-G submetrics for the 65,535 SNPs panel dataset. Each barplot shows **A)** Proximity leakage components: proximity quantile and adversarial distance signals. **B)** Kinship replay components: replay, internal kinship excess, micro haplotype collision, and spectral anomaly. **C)** Trait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

We next examined component-level PRISM-G scores to assess how proximity, relatedness, and trait-linked leakage contribute to overall privacy risk across generators. Figure 3 shows the radar plots of the component scores for each generator. For the 10,000 SNPs panel, GAN and RBM displayed moderate proximity leakage $(0.30 < \mathrm{PLI} < 0.35)$, with kinship replay moderate for GAN $(\mathrm{KRI} = 0.44)$ and low for RBM $(\mathrm{KRI} = 0.02)$. Trait-linked leakage was low for GAN $(\mathrm{TLI} = 0.27)$ but high for RBM $(\mathrm{TLI} = 0.79)$. Genomator presented high proximity leakage under tight constraints that decreased as constraints were relaxed $(0.73 < \mathrm{PLI} < 0.83)$, while kinship replay index remained elevated $(0.75 < \mathrm{KRI} < 0.81)$ and trait-linked leakage stayed low $(0.02 < \mathrm{TLI} < 0.04)$.

On the 65,535 SNPs panel, GAN and RBM showed negligible proximity leakage. RBM displayed negligible kinship replay, whereas GAN showed moderate relatedness ($\mathrm{KRI} = 0.54$). Trait-linked leakage remained minimal for GAN ($\mathrm{TLI} = 0.003$) but elevated for RBM ($\mathrm{TLI} = 0.56$). Genomator's profile shifted toward lower proximity leakage with increasing constraint parameters ($0.05 < \mathrm{PLI} < 0.43$), with intermediate kinship-related contributions ($0.44 < \mathrm{KRI} < 0.56$) and low trait-linked leakage ($0.05 < \mathrm{TLI} < 0.08$). Taken together, the component-level results show that privacy risk can emerge from different underlying mechanisms, including proximity over-closeness, relatedness inflation, and trait-linked exposure via rare genetic variants. Full component estimates are provided in Supplementary Tables 7-10.

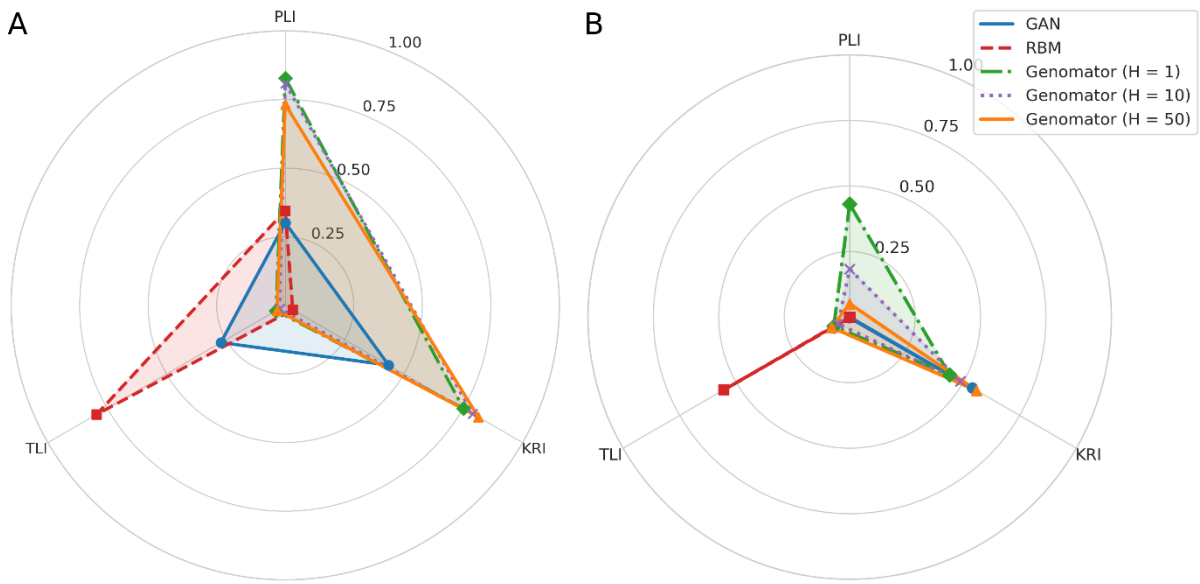

Figure 3. PRISM-G components. Each radar plot shows **A)** Metrics for the 10,000 SNPs panel dataset. **B)** Metrics for the 65,535 SNPs panel dataset.

PRISM-G scores and ranking assessment

We next summarized the component-level privacy risk into a single PRISM-G score to enable compact comparison and ranking of synthetic genome data generators. To this end, we computed mean PRISM-

G scores using bootstrapping (N = 1,000), calibrating each generator against safe and leaky reference datasets to place results on a common risk scale (see Methods). Figure 4 and Supplementary Tables 11-12 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 ($\alpha = 0.08$) and the leaky copycat at 100 ($\beta = 0.75$). Among the learned models, GAN scored the lowest ($\text{PRISM-G} = 35.31$), while RBM scored the highest ($\text{PRISM-G} = 70.37$). Genomator yielded moderate scores on this panel, clustering between 41.58 and 45.53 across constraint settings, consistent with the elevated proximity and kinship leakage signals observed in the submetric components. Based on these scores, the generated datasets from GAN and Genomator were classified as **green** (safest; $\text{PRISM-G} < 50$), whereas the RBM-generated dataset was classified as **amber** (leaky; $50 < \text{PRISM-G} < 90$) relative to the reference anchors.

On the 65,535 SNPs dataset from Chromosome 1, we observed the same relative ordering of generators (Fig. 4b). GAN again showed the lowest risk ($\text{PRISM-G} = 2.10$), while RBM yielded the highest score among the models ($\text{PRISM-G} = 42.76$). Genomator's scores decreased with looser constraints, spanning 10.11 to 26.53 across H settings. Using the same reference anchors as above, overall privacy risk scores on the denser panel were lower than those observed for Chromosome 15. In this setting, RBM remained the highest-risk generator, Genomator's scores tracked its constraint H parameter, and GAN consistently showed low privacy risk. As a result, the datasets were labeled safe ($\text{PRISM-G} < 50$), at higher number of SNPs.

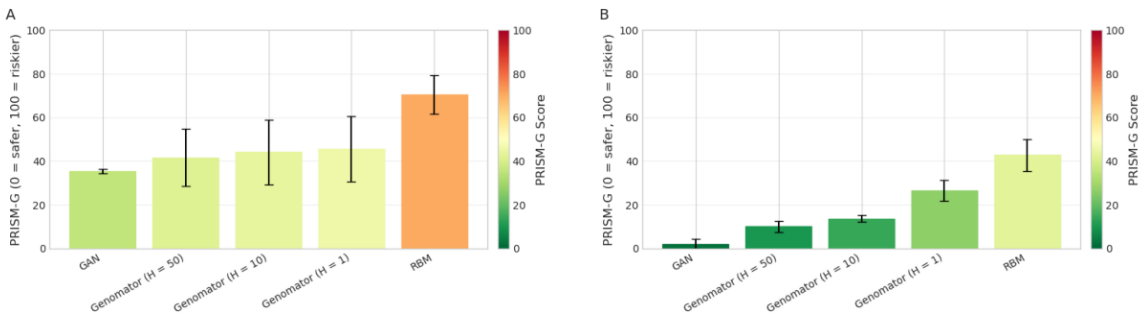

Figure 4. Mean PRISM-G scores for synthetic data generators with $\varepsilon = 0.02$, $\lambda = 0.75$, and $\gamma = 10^{-3}$ as model hyperparameters. **A)** 10,000 SNPs panel datasets. **B)** 65,535 panel datasets.

Using the bootstrap replicates, we then estimated Kendall's rank correlation coefficient to assess how consistently the PRISM-G scores reproduced the model ranking within each dataset. Table 2 shows the summary statistics of the ranking stability test. Given the baseline ordering (RBM > Genomator > GAN; first bootstrap draw), the 65,535 SNPs dataset showed very high correlation ($\bar{\tau} = 0.962$; P-value < 0.018), whereas the 10,000 SNPs dataset showed moderate, non-significant correlation ($\bar{\tau} = 0.699$; P-value < 0.074).

Dataset	$\bar{\tau}$	**SD**	**LCI (95%)**	**UCI (95%)**	**P-value**
10,000 SNPs (Chromosome 15)	0.699	0.287	0.200	1.000	0.074
65,535 SNPs (Chromosome 1)	0.962	0.081	0.800	1.000	0.018

Table 2. Summary statistics of the Kendall ranking stability using $\varepsilon = 0.02$, $\lambda = 0.75$, and $\gamma = 10^{-3}$ as model hyperparameters. SD: standard deviation, LCI: lower confidence interval, UCI: upper confidence interval.

Sensitivity analysis and robustness evaluation

We evaluated the robustness of PRISM-G scores by performing a sensitivity analysis across a range of plausible calibration and regularization settings. In our previous results, PRISM-G was computed using a predefined set of hyperparameters ε, λ, and γ, which control anchor separation and regularization in the noisy-OR aggregation. We recomputed PRISM-G over a grid $\varepsilon \in \{0.02, 0.04, 0.06, 0.08, 0.10\}$, $\lambda \in \{0.75, 0.80, 0.85, 0.90\}$, and $\gamma \in \{10^{-4}, 10^{-3}, 10^{-2}\}$ at 1,000 bootstrap replicates per setting. Figure 5 shows the resulting mean aggregated scores. Across both SNP panels, PRISM-G decreased monotonically with increasing λ, while changes in ε primarily shifted score levels without altering relative ordering. The effect of ridge regularization varied with marker density, modulating the dispersion of scores without altering relative ordering. Overall, rankings and qualitative conclusions remained consistent across the grid, indicating that PRISM-G is robust to plausible variations in calibration and regularization parameters.

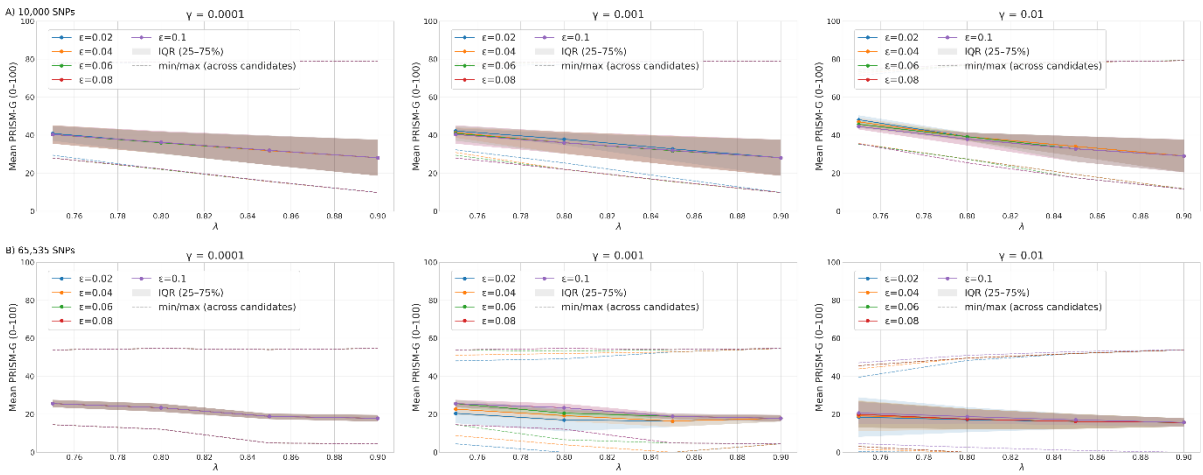

Figure 5. Sensitivity analysis plot to assess the estimated effect of model hyperparameters on PRISM-G scores. Mean PRISM-G scores are shown as a function of the calibration parameter λ for different values of ε. Shaded regions indicate the interquartile range of PRISM-G scores across bootstrap replicates, while dashed horizontal lines denote the minimum and maximum scores observed across

replicates and hyperparameter settings. Panels show results for **a)** 10,000 SNPs panel datasets. **b)** 65,535 SNPs panel datasets. From left to right the plots correspond to $\gamma \in \{10^{-4}, 10^{-3}, 10^{-2}\}$.

We performed a grid search evaluation to select the hyperparameters that maximizes $\bar{\tau}$. Table 3 and Supplementary Tables 13-16 shows the summary results and pairwise differences obtained for both datasets. With the best parameters obtained, the PRISM-G scores lead to a significant ranking with high Kendall's correlation coefficients ($\bar{\tau} > 0.80$; P-value < 0.018). Figure 6 shows the final estimated PRISM-G with significant pairwise comparisons across models. In both datasets, the same ranking was obtained despite differences in PRISM-G estimates.

Dataset	$\boldsymbol{\varepsilon}$	$\boldsymbol{\lambda}$	$\boldsymbol{\gamma}$	$\bar{\boldsymbol{\tau}}$	**SD**	**LCI (95%)**	**UCI (95%)**	**P-value**
10,000 SNPs (Chromosome 15)	0.04	0.75	0.01	0.893	0.219	0.400	1.000	0.018
65,535 SNPs (Chromosome 1)	0.04	0.75	0.01	0.985	0.052	0.800	1.000	0.018

Table 3. Summary statistics of the Kendall ranking stability using the optimal model hyperparameters obtained from grid search. SD: standard deviation, LCI: lower confidence interval, UCI: upper confidence interval.

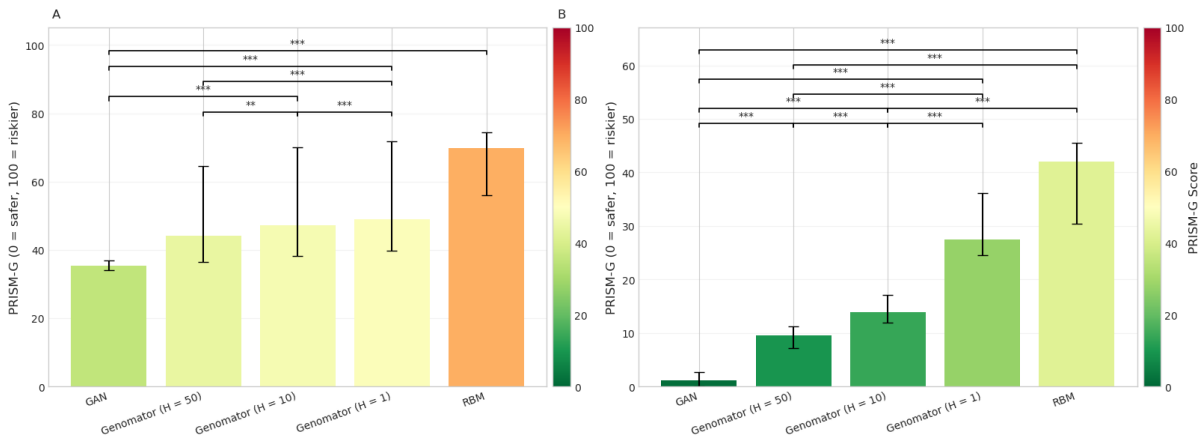

Figure 6. Mean PRISM-G scores for synthetic data generators and pairwise comparisons using the optimal model hyperparameters obtained using grid search. **A)** 10,000 SNPs panel datasets. **B)** 65,535 panel datasets.

Privacy-utility analysis

Using the estimated PRISM-G scores as privacy metrics, we constructed a privacy-utility frontier to evaluate downstream analytical performance while controlling privacy risk. As a representative task, we assessed genetic ancestry prediction using population structure analysis as described in the Evaluation procedure section. Figure 7 shows the performance metrics obtained for both Chromosome datasets. Across generators, synthetic datasets preserved population structure and achieved high utility for ancestry prediction (Performance metric > 90%). Results on the PCA analysis for the synthetic datasets are shown in Supplementary Figures.

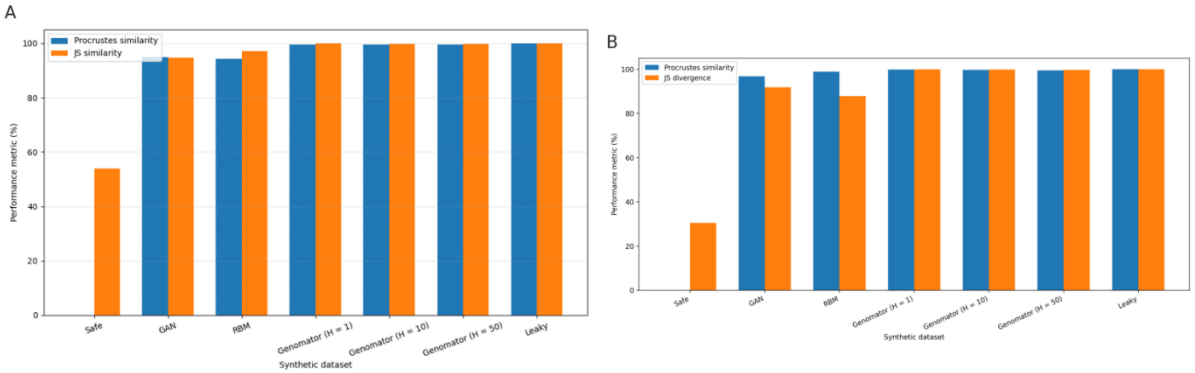

Figure 7. Performance metrics for population structure preservation and ancestry inference utility. **A)** 10,000 SNPs panel datasets. **B)** 65,535 panel datasets.

Pairing the PRISM-G scores with the average performance metrics as a utility measure for population structure analysis yields the privacy-utility frontier shown in Figure 8. The Pareto front represents the set of non-dominated solutions, where improving privacy would require sacrificing utility and vice versa. Genomator configurations lie on this frontier, maintaining high utility while reducing privacy risk as the Hamming constraint increases. GAN also achieves high utility with moderate privacy risks. In contrast, RBM is dominated, exhibiting higher privacy risk without a corresponding utility advantage.

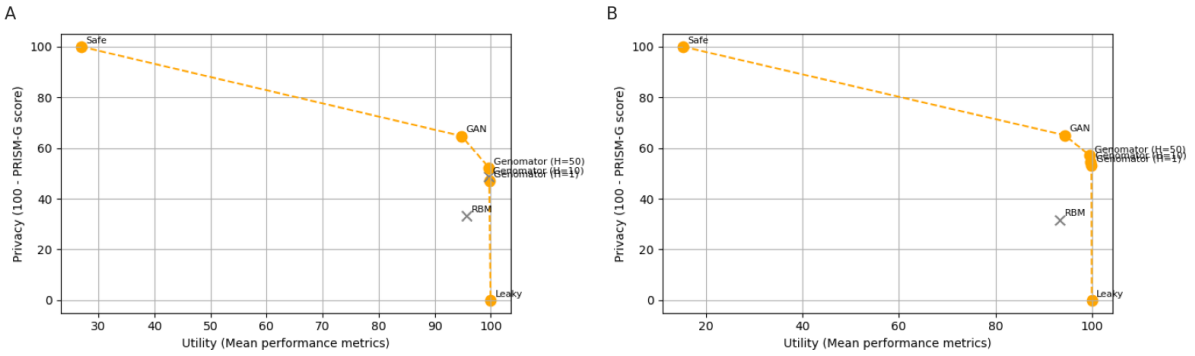

Figure 8. Privacy-utility trade-off curves. **A)** 10,000 SNPs panel datasets. **B)** 65,535 panel datasets.

Discussion

In this study, we introduced PRISM-G, an integrated, model-agnostic framework for assessing privacy risk 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). Central to PRISM-G is the view that privacy risk in synthetic genomic data reflects potential leakage of personal information, which may arise not only from direct resemblance between synthetic and real genomes, but also from preserved relationships among individuals and from genetic signals linked to distinctive traits or characteristics.

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 [30,44]. Moreover, genomic data encode personal information at multiple levels. Distance alone does not establish resistance to concrete attacks or capture risk arising from linked information [45]: long-range familiar search in genealogy resources can re-identify individuals [46] and functional-genomics releases can leak attributes through cryptic channels and data linkage [47,48]. PRISM-G makes distinct leakage pathways explicit, revealing failure modes that single-metric approaches may overlook. 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-range familial search. The trait component (TLI) targets exposure through rare-variant collisions and simple membership-inference advantages. These components are recognized vectors for identification or singling-out [38]. 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 [49,50].

Applying PRISM-G to multiple synthetic generators on two marker panels from the 1KGP dataset showed that privacy risk concentrates along different axes depending on model design and

parametrization, as highlighted with prior attack-based evaluation of synthetic genomic data. Oprisanu and colleagues 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 [27]. Using the GAN and RBM baselines from Yelmen et al. and a logic-based generator (Genomator) [20,21], PRISM-G component scores and submetrics reveal distinct patterns that reflect different modes of potential personal information leakage: 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 that can produce synthetic data with limited diversity and, hence, hamper their effectiveness [51]. However, methods and settings that improve coverage can mitigate collapse and increase sample diversity [20,40]. 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, amplifying trait-linked and familial leakage signals, thus elevating the corresponding PRISM-G components. Unintended memorization is a documented risk in generative models and can disproportionately affect uncommon samples, suggesting a plausible mechanism for our empirical findings [52,53]. 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.

Beyond interpreting PRISM-G and its components (PLI, KRI, TLI), Kendall's rank stability test offers a robust way to compare models, indicating which one tends to produce safer, leaky, or risky datasets. Across both maker panels and all models evaluated, our results suggest that, for the 1KGP data, GAN

produced the safest synthetic genomic dataset in PRISM-G terms, whereas RBM produced the riskiest. Taken together, our results reinforce that genomic "privacy" is inherently multi-faceted rather than a single-number attribute, with each model exhibiting distinct leakage patterns across genomic contexts that extend beyond simple distance metrics or rare-variant content.

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 and statistical ranking 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 TLI). 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.**

Apart from privacy exposure diagnostics, PRISM-G also enables joint evaluation of privacy and analytical performance through a privacy-utility frontier. Applying this framework shows that although all generators achieved similarly high utility for ancestry prediction their privacy risks differed substantially. Genomator configurations occupied favorable regions of the frontier, maintaining high utility while progressively reducing privacy risk as the Hamming distance constrained increased, whereas RBM exhibited higher privacy exposure without a corresponding gain in analytical performance. These results indicate that generators with comparable analytical utility may differ markedly in privacy risk.

Interpretable, component-level privacy auditing is essential for governance in multi-skateholder settings where trust depends on demonstrable safeguards. In European cross-border contexts, decisions about data sharing must be justified under heterogeneous legal interpretations and risk tolerances, making a single opaque similarity score insufficient [10,12]. PRISM-G addresses this need by decomposing privacy exposure into proximity, kinship replay, and trait-linked (rare variant) mechanisms, enabling systematic auditing of synthetic genomic data and transparent comparison across generators. By making these mechanisms explicit, PRISM-G supports structured assessment of privacy risk and provides a basis for evaluating privacy-utility trade-offs for specific genomic tasks. In addition, the framework can support equity-aware evaluation, as proximity leakage related to ancestry structure and exposure driven by rare-variant uniqueness can vary across populations and may warrant additional scrutiny in future applications.

There are several limitations to note. PRISM-G does not provide a formal privacy guarantee (e.g., differential privacy). It is an empirical 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, this work suggests several directions. Methodologically, refining the algorithms 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. Finally, frameworks such as PRISM-G could support broader auditing and governance workflows to provide structured evaluation of privacy risk, anonymization strategies, and data-sharing practices for future use of synthetic genomic data.

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.

Declaration of generative AI and AI-assisted technologies in the writing process

The authors used ChatGPT to assist with reviewing the manuscript for grammar, clarity, and consistency, and to support debugging of code. Following the use of this tool, the reviewers and edited all generated outputs. The authors take full responsibility for the contents presented in this publication.

Data availability

The 1000 Genomes Project is an open-access dataset available at https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/. All synthetic genomes used in this study have been deposited in the PRISM-G repository on GitHub at https://github.com/alejocrojo09/prismg/tree/master/example_data.

Code availability

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

References

1. Venturini P, Faria PL, Cordeiro JV. AI and omics technologies in biobanking: Applications and challenges for public health. *Public Health* 2025;**243**:105726. https://doi.org/10.1016/j.puhe.2025.105726.

2. Chen YM, Hsiao TH, Lin CH *et al.* Unlocking precision medicine: clinical applications of integrating health records, genetics, and immunology through artificial intelligence. *J Biomed Sci* 2025;**32**(1):16. https://doi.org/10.1186/s12929-024-01110-w.

3. Mani S, Lalani SR, Pammi M. Genomics and multiomics in the age of precision medicine. *Pediatr Res* 2025;**97**(4):1399–410. https://doi.org/10.1038/s41390-025-04021-0.

4. Conduah AK, Ofoe S, Siaw-Marfo D. Data privacy in healthcare: Global challenges and solutions. *Digit Health* 2025;**11**:20552076251343959. https://doi.org/10.1177/20552076251343959.

5. Evans BJ, Jarvik GP. Impact of HIPAA's minimum necessary standard on genomic data sharing. *Genet Med* 2018;**20**(5):531–5. https://doi.org/10.1038/gim.2017.141.

6. Shabani M, Marelli L. *Re-identifiability of Genomic Data and the GDPR*. n.d.

7. Gymrek M, McGuire AL, Golan D *et al.* Identifying Personal Genomes by Surname Inference. *Science* 2013;**339**(6117):321–4. https://doi.org/10.1126/science.1229566.

8. Bonomi L, Huang Y, Ohno-Machado L. Privacy challenges and research opportunities for genomic data sharing. *Nat Genet* 2020;**52**(7):646–54. https://doi.org/10.1038/s41588-020-0651-0.

9. Stark Z, Glazer D, Hofmann O *et al.* A call to action to scale up research and clinical genomic data sharing. *Nat Rev Genet* 2025;**26**(2):141–7. https://doi.org/10.1038/s41576-024-00776-0.

10. Crum E, Taelman R, Buelens B *et al.* Genomic sequence data sharing for clinical practice: A scoping review. *Comput Biol Med* 2026;**200**:111335. https://doi.org/10.1016/j.compbiomed.2025.111335.

11. Middleton A, Milne R, Almarri MA *et al.* Global Public Perceptions of Genomic Data Sharing: What Shapes the Willingness to Donate DNA and Health Data? *Am J Hum Genet* 2020;**107**(4):743–52. https://doi.org/10.1016/j.ajhg.2020.08.023.

12. Royo R, Garcia E, Schlünder I *et al.* Genomic data sharing in research across Europe: legal challenges and upcoming opportunities within the European Health Data Space. *Eur J Public Health* 2025;**35**(Supplement_3):iii25–31. https://doi.org/10.1093/eurpub/ckaf070.

13. Giuffrè M, Shung DL. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *Npj Digit Med* 2023;**6**(1):186. https://doi.org/10.1038/s41746-023-00927-3.

14. Baumdicker F, Bisschop G, Goldstein D *et al.* Efficient ancestry and mutation simulation with msprime 1.0. *Genetics* 2022;**220**(3):iyab229. https://doi.org/10.1093/genetics/iyab229.

15. Mendes JM, Barbar A, Refaie M. Synthetic data generation: a privacy-preserving approach to accelerate rare disease research. *Front Digit Health* 2025;**7**:1563991. https://doi.org/10.3389/fdgth.2025.1563991.

16. Wharrie S, Yang Z, Raj V *et al.* HAPNEST: efficient, large-scale generation and evaluation of synthetic datasets for genotypes and phenotypes. *Bioinformatics* 2023;**39**(9):btad535. https://doi.org/10.1093/bioinformatics/btad535.

17. Holt KE, Inouye M. Genomic data sharing: you don't know what you've got (till it's gone). *Nat Rev Genet* 2025;**26**(5):293–4. https://doi.org/10.1038/s41576-025-00820-7.

18. Oliva A, Kaphle A, Reguant R *et al.* Future-proofing genomic data and consent management: a comprehensive review of technology innovations. *GigaScience* 2024;**13**:giae021. https://doi.org/10.1093/gigascience/giae021.

19. Kenneweg P, Dandinasivara R, Luo X *et al.* Generating synthetic genotypes using diffusion models. *Bioinformatics* 2025;**41**(Supplement_1):i484–92. https://doi.org/10.1093/bioinformatics/btaf209.

20. Yelmen B, Decelle A, Boulos LL *et al.* Deep convolutional and conditional neural networks for large-scale genomic data generation. *PLOS Comput Biol* 2023;**19**(10):e1011584. https://doi.org/10.1371/journal.pcbi.1011584.

21. Burgess MA, Hosking B, Reguant R *et al.* Privacy-hardened and hallucination-resistant synthetic data generation with logic-solvers. *Bioinformatics* 2025;**41**(12):btaf600. https://doi.org/10.1093/bioinformatics/btaf600.

22. Battey CJ, Coffing GC, Kern AD. Visualizing population structure with variational autoencoders. *G3 GenesGenomesGenetics* 2021;**11**(1):jkaa036. https://doi.org/10.1093/g3journal/jkaa036.

23. Wang X, Min S, Vaidya J. *Exploring the Use of Artificial Genomes for Genome-Wide Association Studies through the Lens of Utility and Privacy*. n.d.

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

25. Yelmen B, Decelle A, Ongaro L *et al.* Creating artificial human genomes using generative neural networks. *PLOS Genet* 2021;**17**(2):e1009303. https://doi.org/10.1371/journal.pgen.1009303.

26. Shokri R, Stronati M, Song C *et al.* Membership Inference Attacks against Machine Learning Models, arXiv:1610.05820. Preprint, arXiv, 31 Mar. 2017. https://doi.org/10.48550/arXiv.1610.05820.

27. Oprisanu B, Ganev G, Cristofaro ED. On Utility and Privacy in Synthetic Genomic Data, arXiv:2102.03314. Preprint, arXiv, 18 Jan. 2022. https://doi.org/10.48550/arXiv.2102.03314.

28. Gürsoy G, Li T, Liu S *et al.* Functional genomics data: privacy risk assessment and technological mitigation. *Nat Rev Genet* 2022;**23**(4):245–58. https://doi.org/10.1038/s41576-021-00428-7.

29. Thomas M, Mackes N, Preuss-Dodhy A *et al.* Assessing Privacy Vulnerabilities in Genetic Data Sets: Scoping Review. *JMIR Bioinforma Biotechnol* 2024;**5**:e54332. https://doi.org/10.2196/54332.

30. Kaabachi B, Despraz J, Meurers T *et al.* A scoping review of privacy and utility metrics in medical synthetic data. *Npj Digit Med* 2025;**8**(1):60. https://doi.org/10.1038/s41746-024-01359-3.

31. Kuru T. Genetic Data: The Achilles' Heel of the GDPR? *Eur Data Prot Law Rev* 2021;**7**(1):45–58. https://doi.org/10.21552/edpl/2021/1/8.

32. Patterson N, Price AL, Reich D. Population Structure and Eigenanalysis. *PLoS Genet* 2006;**2**(12):e190. https://doi.org/10.1371/journal.pgen.0020190.

33. Manichaikul A, Mychaleckyj JC, Rich SS *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* 2010;**26**(22):2867–73. https://doi.org/10.1093/bioinformatics/btq559.

34. Madsen BE, Browning SR. A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. *PLoS Genet* 2009;**5**(2):e1000384. https://doi.org/10.1371/journal.pgen.1000384.

35. The 1000 Genomes Project Consortium, Corresponding authors, Auton A *et al.* A global reference for human genetic variation. *Nature* 2015;**526**(7571):68–74. https://doi.org/10.1038/nature15393.

36. Goudet J, Kay T, Weir BS. How to estimate kinship. *Mol Ecol* 2018;**27**(20):4121–35. https://doi.org/10.1111/mec.14833.

37. VanRaden PM. Efficient Methods to Compute Genomic Predictions. *J Dairy Sci* 2008;**91**(11):4414–23. https://doi.org/10.3168/jds.2007-0980.

38. Shringarpure SS, Bustamante CD. Privacy Risks from Genomic Data-Sharing Beacons. *Am J Hum Genet* 2015;**97**(5):631–46. https://doi.org/10.1016/j.ajhg.2015.09.010.

39. Yale A, Dash S, Dutta R *et al. Privacy Preserving Synthetic Health Data*. n.d.

40. Kuo NIH, Garcia F, Sönnerborg A *et al.* Generating synthetic clinical data that capture class imbalanced distributions with generative adversarial networks: Example using antiretroviral therapy for HIV. *J Biomed Inform* 2023;**144**:104436. https://doi.org/10.1016/j.jbi.2023.104436.

41. Oldoni F, Kidd KK, Podini D. Microhaplotypes in forensic genetics. *Forensic Sci Int Genet* 2019;**38**:54–69. https://doi.org/10.1016/j.fsigen.2018.09.009.

42. Seidman DN, Shenoy SA, Kim M *et al.* Rapid, Phase-free Detection of Long Identity-by-Descent Segments Enables Effective Relationship Classification. *Am J Hum Genet* 2020;**106**(4):453–66. https://doi.org/10.1016/j.ajhg.2020.02.012.

43. Burgess MA, Hosking B, Reguant R *et al. Privacy-Hardened and Hallucination-Resistant Synthetic Data Generation with Logic-Solvers*. n.d.

44. Ganev G. Synthetic Data, Similarity-based Privacy Metrics, and Regulatory (Non-)Compliance, arXiv:2407.16929. Preprint, arXiv, 26 July 2024. https://doi.org/10.48550/arXiv.2407.16929.

45. Zhang Z, Yan C, Malin BA. Membership inference attacks against synthetic health data. *J Biomed Inform* 2022;**125**:103977. https://doi.org/10.1016/j.jbi.2021.103977.

46. Edge MD, Coop G. Attacks on genetic privacy via uploads to genealogical databases. *eLife* 2020;**9**:e51810. https://doi.org/10.7554/eLife.51810.

47. Gürsoy G, Brannon CM, Navarro FCP *et al.* FANCY: fast estimation of privacy risk in functional genomics data. *Bioinformatics* 2021;**36**(21):5145–50. https://doi.org/10.1093/bioinformatics/btaa661.

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

49. Pilgram L, Dankar FK, Drechsler J *et al.* A consensus privacy metrics framework for synthetic data. *Patterns* July 2025:101320. https://doi.org/10.1016/j.patter.2025.101320.

50. Zamzmi G, Subbaswamy A, Sizikova E *et al.* Scorecard for synthetic medical data evaluation. *Commun Eng* 2025;**4**(1):130. https://doi.org/10.1038/s44172-025-00450-1.

51. Goodfellow I. NIPS 2016 Tutorial: Generative Adversarial Networks, arXiv:1701.00160. Preprint, arXiv, 3 Apr. 2017. https://doi.org/10.48550/arXiv.1701.00160.

52. Carlini N, Liu C, Erlingsson Ú *et al.* The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks, arXiv:1802.08232. Preprint, arXiv, 16 July 2019. https://doi.org/10.48550/arXiv.1802.08232.

53. Chen Y, Wang S, Zou D *et al.* SIDE: Surrogate Conditional Data Extraction from Diffusion Models, arXiv:2410.02467. Preprint, arXiv, 1 Aug. 2025. https://doi.org/10.48550/arXiv.2410.02467.