Bioinformatics, 2026, **42**, btag377
https://doi.org/10.1093/bioinformatics/btag377
25th European Conference on Computational Biology supplement

OXFORD

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]Environmental Intelligence (EI), Flemish Institute for Technological Research (VITO), Mol, 2400, Belgium
*Corresponding authors. Alejandro Correa Rojo, ESAT-STADIUS, KU Leuven, Leuven, 3001, Belgium. Email: alejandro.correarojo@esat.kuleuven.be; Gökhan Ertaylan, Environmental Intelligence (EI), Flemish Institute for Technological Research (VITO), Mol, 2400, Belgium. Email: gokhan.ertaylan@vito.be.

Abstract

Motivation: **Synthetic genomic data promises broader data access, but unresolved privacy risks remain a major concern. Existing evaluations often rely on similarity-based metrics that measure proximity between real and synthetic genomes, overlooking additional mechanisms through which genomic information may leak.**

Results: **We introduce PRISM-G, a model-agnostic framework that quantifies privacy exposure in synthetic genomic data across three complementary components: proximity to real genomes in genetic-coordinate space, replay of familial or population-structure patterns, and trait-linked exposure through 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. By pairing PRISM-G with downstream utility metrics, the framework also enables analysis of privacy–utility trade-offs across generative models. 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 arise along different axes across models and marker densities, demonstrating that a single similarity-based metric is insufficient to characterize genomic privacy risk.**

Availability and implementation: **The source code of PRISM-G is available at https://github.com/alejocrojo09/prismg.**

Introduction

The increasing availability of population-scale genomic resources has transformed biomedical research by enabling systematic characterization of human genetic variation and its role in health and disease (Venturini *et al.* 2025). Large international sequencing initiatives and national biobanks now host datasets comprising large collections of human genomes, supporting applications ranging from rare disease gene discovery to population-scale studies of complex traits and the development of precision medicine (Chen *et al.* 2025, Mani *et al.* 2025). However, the growing availability of genomic data also raises persistent concerns regarding privacy and responsible data sharing. Human genomes encode highly sensitive information and remain inherently identifiable even when explicit identifiers are removed, creating risks of re-identification or attribute disclosure (Conduah *et al.* 2025, Evans and Jarvik 2018, Shabani and Marelli 2019, Gymrek *et al.* 2013).

These concerns are particularly relevant in collaborative research environments where genomic data must be shared across institutions and national borders (Bonomi *et al.* 2020, Stark *et al.* 2025). In Europe, large-scale genomic research infrastructures operate across multiple jurisdictions, where differences in governance frameworks and national interpretations of data protection obligations complicate cross-border data sharing (Crum *et al.* 2026). Although initiatives such as the **European Health Data Space (EHDS)** (https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space-regulation-ehds_en) and the **European Genomic Data Infrastructure (GDI)** (https://gdi.onemilliongenomes.eu/) aim to facilitate the secure secondary use of health and genomic data across member states, practical data sharing often remains constrained by disagreements over acceptable privacy risks and by the lack of widely accepted technical evidence demonstrating that shared genomic data cannot be linked back to individuals (Royo *et al.* 2025).

In response to these challenges, synthetic data generation has emerged as a strategy to expand data accessibility while reducing privacy risks and preserving analytical utility without direct reliance on real individual-level records (Giuffrè and Shung 2023). In genetics, several generative approaches, including haplotype resampling (Wharrie *et al.* 2023), deep learning architectures (Yelmen *et al.* 2021, 2023, Kenneweg *et al.* 2025), and logic-based methods (Burgess *et al.* 2025), have been developed to produce artificial genotype datasets that reproduce key genomic properties such as allele frequencies, linkage disequilibrium (LD), and population

© The Author(s) 2026. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from academic.oup.com/bioinformatics/article/42/Supplement_2/btag377/8767287 by INACTIVE user on 29 September 2026

structure. These synthetic datasets can support exploratory analyses, method development, and collaborative research where access to individual genomic data is restricted, while enabling applications in rare disease research, population genetics, and genetic risk estimation (Mendes *et al.* 2025).

Despite these advances, evaluating the privacy risk of synthetic genomic datasets remains challenging. Many studies rely on distance-based similarity measurements to estimate how closely synthetic genomes resemble individuals in the original dataset. Typical approaches include nearest-neighbor or Hamming distance comparisons between real and synthetic genomes, and low similarity is often interpreted as evidence that synthetic datasets are privacy-preserving (Yelmen *et al.* 2021, 2023, Wharrie *et al.* 2023, Burgess *et al.* 2025, Kenneweg *et al.* 2025). However, this assumption can be misleading. Previous studies have shown that Membership Inference Attacks (MIAs) can determine whether a specific individual was used to train a generative model even when synthetic genomes exhibit low similarity to the training data (Oprisanu *et al.* 2022).

These limitations arise because genomic information can reveal identity through mechanisms beyond simple genome similarity. Synthetic datasets may reproduce relatedness patterns from the training data, enabling familial or genealogical matching (Edge and Coop, 2020). Distinctive features, such as rare variants or unusual mutation combinations, may uniquely characterize individuals or small related groups (Thomas *et al.* 2024). Personal information may also be inferred through linkage with external resources, including functional genomics data (Gürsoy *et al.* 2022). As a result, datasets that appear safe under similarity-based evaluations may still retain signals enabling indirect identification. These observations suggest that privacy exposure cannot be captured by a single metric; rather, risk arises across complementary representations of genetic information, such as population structure, relatedness, and distinctive variant patterns, requiring evaluation frameworks that assess these signals jointly.

Building on this perspective, we introduce **PRISM-G (Privacy Risk Integrated Score for Multi-representation Genomes)**, a framework for systematically assessing privacy exposure in synthetic genomic datasets. PRISM-G evaluates privacy risk across complementary genomic representations that capture proximity between real and synthetic genomes, patterns of genetic relatedness, and trait-linked genetic features, producing a privacy score that enables consistent comparison across synthetic datasets and generative models. Beyond providing a compact privacy metric, the framework generates interpretable diagnostics that reveal the mechanisms underlying potential privacy leakage.

To illustrate this approach, we applied PRISM-G to synthetic cohorts generated from multiple generative model families, including generative adversarial networks (GANs), restricted Boltzmann machines (RBMs), and a logic-based SAT solver, using data subsets from the 1000 Genomes Project (1000 Genomes Project Consortium 2015).

PRISM-G framework: algorithmic background and overview

Genomic data can be represented through complementary statistical structures that capture population variation and relatedness. Common representations include coordinate embeddings derived from dimensionality reduction, relational representations based on genetic relationship matrices (GRMs), and feature-level signals such as rare genetic variants associated with individual characteristics (Patterson *et al.* 2006, Madsen and Browning 2009, Goudet *et al.* 2018). These representations are widely used in population genetics to quantify genetic similarity, infer relatedness, and characterize variant distributions across individuals.

Building on these genomic representations, we developed PRISM-G, which evaluates privacy exposure by comparing real genomes with synthetic cohorts across three complementary components: the **Proximity Leakage Index (PLI)**, measuring similarity between real and synthetic genomes; the **Kinship Replay Index (KRI)**, capturing replication of relatedness patterns such as kinship and haplotype sharing; and the **Trait-Linked Leakage Index (TLI)**, detecting preservation of rare-variant signals. Each component reflects a distinct representation of genomic variation and is decomposed into interpretable submetrics that diagnose specific leakage mechanisms. A structured comparison of the algorithmic foundations of these components and their relation to existing genomic and privacy metrics is provided in the Supplementary Data.

The component metrics are subsequently aggregated into a single composite score summarizing the overall privacy risk of a synthetic dataset. This unified score enables statistical comparison of datasets generated by different models and provides a quantitative measure of privacy exposure. When combined with downstream utility measurements, the PRISM-G score can also be used to construct privacy–utility frontiers that characterize trade-offs between analytical utility and privacy risk. An overview of the PRISM-G workflow is illustrated in Fig. 1.

In the following sections we formally define each component metric and the aggregation procedure used to derive the PRISM-G score, while the full mathematical formulation and methodological details are provided in the Supplementary Data.

Component metrics

Proximity Leakage Index

The PLI metric evaluates whether synthetic genomes lie unusually close to real individuals in genetic coordinate space. Genotypes are first standardized by allele frequencies and embedded using principal component analysis (PCA), producing a low-dimensional representation of population structure (Patterson *et al.* 2006). Distances between real and synthetic samples are then computed using nearest-neighbor Euclidean distance in the principal component space. From these distances, PLI combines two complementary submetrics:

- **Quantile proximity ratio (r_p):** evaluates whether the smallest distances between synthetic and real individuals fall below the expected distribution of distances among real individuals.
- **Adversarial proximity (r_A):** measures whether holdout individuals are closer to synthetic genomes than to other real individuals.

Together, these diagnostics detect identity proximity, where synthetic samples may be sufficiently similar to real genomes to

Downloaded from academic.oup.com/bioinformatics/article/42/Supplement_2/btag377/8767287 by INACTIVE user on 29 September 2026

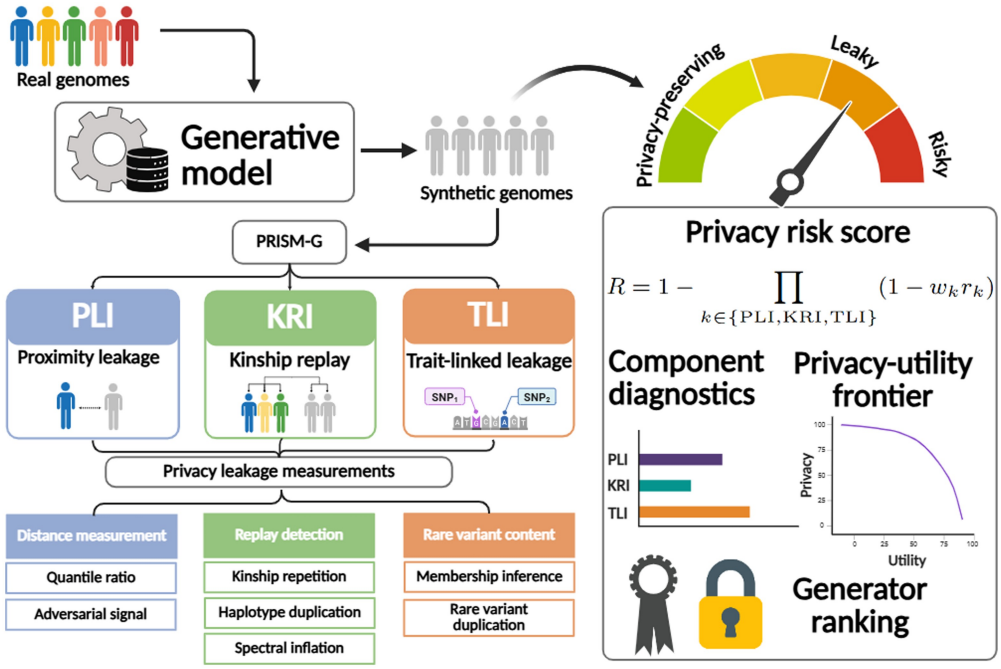


Figure 1 Overview of the PRISM-G framework. Imaged generated with BioRender.

enable singling-out or linkage with external genetic databases. The overall score is defined as

$$\mathrm{PLI} = \max(r_p, r_A). \tag{1}$$

Kinship Replay Index

The KRI metric captures privacy risks arising from the reproduction of relational genetic structure. Genotypes are used to estimate the GRM following the method proposed by VanRaden, which quantifies pairwise relatedness among individuals (VanRaden 2008). In addition, local haplotype patterns derived from genotype windows are used to identify repeated genotype segments indicative of shared ancestry (Oldoni *et al.* 2019). It aggregates four diagnostics derived from the GRM and haplotype structure:

- **Replay (r_{replay}):** evaluates whether the distribution of close-kin relationships observed in the real dataset is reproduced in the synthetic cohort.
- **Internal kinship excess (r_{IKE}):** measures whether synthetic samples exhibit elevated mutual relatedness relative to the real data by comparing the fraction of kinship values exceeding predefined thresholds.
- **Micro-haplotype collisions (r_{HAP}):** detects repeated short genotype patterns by scanning sliding windows of SNPs and quantifying identical haplotype segments shared across synthetic individuals.
- **Spectral inflation (r_{SPEC}):** identifies concentration of relatedness structure through inflation of the leading eigenvalue of the GRM.

Together these diagnostics capture structural replay, where generative models reproduce genealogical patterns that may enable familial matching or inference of hidden relationships. The overall score is defined as

$$\mathrm{KRI} = \max(r_{\mathrm{replay}}, r_{\mathrm{IKE}}, r_{\mathrm{HAP}}, r_{\mathrm{SPEC}}). \tag{2}$$

Trait-Linked Leakage Index

The TLI metric measures privacy exposure arising from distinctive feature-level signals encoded in rare variants. Rare variants can act as highly identifying genomic features and may reveal sensitive information when preserved in synthetic datasets. TLI therefore evaluates whether rare variant burden patterns and uncommon variant signals observed in real genomes are reproduced in synthetic samples (Madsen and Browning 2009). It combines two signals:

- **Membership inference signal (r_{mia}):** evaluates whether individuals used to train the generative model can be distinguished from holdout individuals using rare variant burden scores and a nearest-neighbor membership classifier.
- **Rare variant uniqueness (r_{uniq}):** detects whether uncommon variants appear more frequently in the synthetic cohort than expected.

Together these signals summarize the extent to which rare variant patterns are preserved in the synthetic dataset. The overall score is defined as

$$\mathrm{TLI} = \max(r_{\mathrm{mia}}, r_{\mathrm{uniq}}). \tag{3}$$

Aggregation of privacy risk

Each component produces a normalized risk score in the interval $[0, 1]$. Because the three components represent distinct

Downloaded from academic.oup.com/bioinformatics/article/42/Supplement_2/btag377/8767287 by INACTIVE user on 29 September 2026

leakage mechanisms, PRISM-G aggregates them using a risk-averse noisy-OR formulation:

$$R_{\text{raw}} = 1 - \prod_{k \in \{\text{PLI,KRI,TLI}\}} (1 - w_k r_k) \quad (4)$$

where r_k denotes the component risk and w_k are non-negative weights constrained to sum to one.

The noisy-OR formulation reflects that privacy exposure may arise through multiple alternative leakage pathways, including proximity, kinship structure, or rare-variant leakage. This differs from the within-component max operator, which conservatively flags any strong submetric-level signal. When average-case rather than worst-case exposure is preferred, PRISM-G can alternatively use mean or median submetric aggregation. Future extensions may also explore parameterized power-mean aggregation to interpolate between average-case and worst-case risk summaries.

Raw privacy scores depend on dataset characteristics such as cohort size, allele-frequency spectrum, and population structure, and therefore cannot be directly compared across experiments. PRISM-G therefore calibrates raw risk values using reference generators that define empirical lower and upper bounds of privacy exposure.

Two anchors are used: a *safe baseline generator*, which preserves allele frequencies while removing correlation structure, and a *leaky reference generator*, which intentionally replicates training samples and reproduce their relational structure. These anchors define calibration parameters α and β, representing the expected raw risk under the safe and leaky references, respectively. Raw risk values are then mapped onto a standardized 0–100 PRISM-G privacy scale:

$$\text{PRISM} - \text{G} = 100 \times \frac{R_{\text{raw}} - \alpha}{\beta - \alpha} \quad (5)$$

To compare privacy risk across generators, PRISM-G evaluates generator rankings using Kendall's rank correlation coefficient. Rankings are recomputed across bootstrap samples and compared with the baseline ranking obtained from the full dataset. The average Kendall correlation measures the agreement between rankings and reflects the stability of the resulting PRISM-G ordering.

The component weights w_k are calibrated using ridge regularization based on the anchor generators. Hyperparameters are explored via grid search, and candidate configurations are evaluated by the Kendall ranking stability. The configuration that maximizes the average Kendall correlation is selected to produce the final PRISM-G scores. The resulting score can be stratified into predefined risk categories (e.g. quartiles) to facilitate interpretation and classify datasets as privacy-preserving, intermediate (leaky), or high-risk.

Because PRISM-G is calibrated using dataset-specific anchors, **absolute scores should not be interpreted as universal privacy probabilities.** Scores are most directly comparable across generators evaluated on the same dataset under the same calibration setting. Comparisons across datasets or cohorts should therefore recalibrate the anchors or report component-level PLI, KRI, and TLI values alongside the calibrated score.

Privacy-utility analysis

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 for privacy protection, PRISM-G can also be paired with task-specific utility metrics to analyze privacy–utility trade-offs, enabling systematic comparison between models and identification of generators that achieve favorable balances between analytical utility and privacy protection.

Evaluation

Experimental setup

We evaluated PRISM-G on synthetic genomic cohorts generated by three model families: generative adversarial networks (GAN), restricted Boltzmann machines (RBM), and Genomator, a SAT-based genome generator (Burgess *et al.* 2025). GAN and RBM datasets were obtained from Yelmen *et al.* (2021), while Genomator cohorts were generated with varying Hamming distance constraints, which control the minimum number of variant differences allowed between synthetic and training samples.

All experiments used the 1000 Genomes Project Phase 3 (1KGP) dataset comprising 2504 individuals. Analyses were conducted on two genotype panels: a 10 000-SNP subset from Chromosome 15 and a 65 535-SNP subset from Chromosome 1. For each experiment, the real dataset was partitioned into training (80%) and holdout (20%) cohorts, and synthetic datasets were generated to match the size of the holdout set.

For calibration of PRISM-G scores, a synthetic dataset generated using a binomial sampler based on 1KGP allele frequencies served as the safe dataset, and a copycat dataset served as the leaky dataset. Component weights were calibrated via ridge regularization against these anchors using a grid search over calibration hyperparameters (Supplementary Data), and ranking stability was evaluated for each configuration using Kendall's rank correlation across bootstrap replicates. This procedure also served as a sensitivity analysis, assessing how hyperparameter choices affect the dispersion and relative ordering of PRISM-G scores across generators.

To characterize population structure, PCA was applied to the real training data, generating embeddings that visualize genetic clusters using 1KGP population labels. Synthetic samples were projected onto the same coordinate space to ensure comparability between real and synthetic genomes. Population-structure preservation was quantified using Procrustes similarity between centroids of real and synthetic ancestry clusters in PCA space and Jensen–Shannon (JS) similarity derived from ancestry label distributions. Downstream predictive performance was evaluated using a train-real, test-synthetic (TRTS) protocol, where a random forest classifier trained on real genomes predicted ancestry labels in the synthetic datasets. We averaged Procrustes similarity and JS similarity as an overall utility metric capturing both structure preservation and predictive performance.

We present the main results for the Chromosome 1 subset (65 535 SNPs) and compare them with findings from the Chromosome

Downloaded from academic.oup.com/bioinformatics/article/42/Supplement_2/btag377/8767287 by INACTIVE user on 29 September 2026

15 dataset (10 000 SNPs). A full description of the Chromosome 15 results is provided in the Supplementary Data.

Privacy leakage diagnostics

We first evaluated PRISM-G submetrics for each synthetic dataset to illustrate how different generators exhibit distinct privacy leakage mechanisms prior to aggregation into PRISM-G scores. Figure 2 summarizes the privacy leakage submetrics for the Chromosome 1 dataset. On this dataset, PLI diagnostics indicated no proximity-based leakage for GAN and RBM (Fig. 2A). In contrast, Genomator showed a proximity signal that depended on the Hamming distance constraint $H \in \{1, 10, 50\}$. Under tight constraints ($H = 1$), synthetic genomes appeared unusually close to holdout individuals, producing the strongest proximity signal. Increasing the constraint progressively reduced this effect, with proximity exposure decreasing at $H = 10$ and becoming minimal at $H = 50$.

Kinship replay diagnostics revealed generator-specific patterns (Fig. 2B). GAN showed moderate spectral inflation ($r_{SPEC} = 0.541$), suggesting partial reproduction of global relatedness structure. In contrast, RBM exhibited negligible kinship leakage across all KRI submetrics. Genomator displayed a constraint-dependent pattern modulated by the Hamming distance parameter H: kinship-related signals increased as the constraint was relaxed, with higher H values showing stronger haplotype-collision and spectral signals (r_{HAP} and r_{SPEC}), while replay signals remained comparatively modest. In the present experiments, r_{HAP} should be interpreted as a local haplotype diagnostic over the assayed SNP panel, rather than a whole-genome identity-by-descent (IBD) estimate. For full-scale deployment, r_{HAP} should be extended to genome-wide segments, as haplotype sharing depends on both marker density and genomic span.

For the TLI component, membership inference signals were generally low across models (Fig. 2C), although small r_{mia} signals were observed for RBM and Genomator. Rare-variant uniqueness showed stronger differences: GAN and Genomator exhibited negligible r_{uniq}, whereas RBM displayed elevated rare-variant uniqueness ($r_{uniq} = 0.555$), indicating stronger preservation of distinctive variant configurations.

Applying the same analysis to the Chromosome 15 dataset yielded similar qualitative patterns (Supplementary Fig. S1), although submetric values were generally higher for this lower-density SNP panel. Genomator again showed decreasing proximity signals with increasing H, kinship metrics were dominated by spectral-relatedness signals for RBM and Genomator, and RBM and GAN exhibited elevated rare-variant uniqueness. Detailed results for both datasets are provided in Supplementary Tables 1–6 in the Supplementary Data.

Component-level PRISM-G scores

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 a radar plot of the component scores for the Chromosome 1 dataset. GAN and RBM showed negligible proximity leakage. GAN showed moderate relatedness (KRI = 0.541), whereas RBM displayed low kinship replay (KRI = 0.002). Trait-linked leakage remained minimal for GAN (TLI = 0.003) but was elevated for RBM (TLI = 0.555). Genomator showed a constraint-dependent profile, with proximity leakage decreasing as the Hamming constraint increased ($0.052 < \text{PLI} < 0.431$),

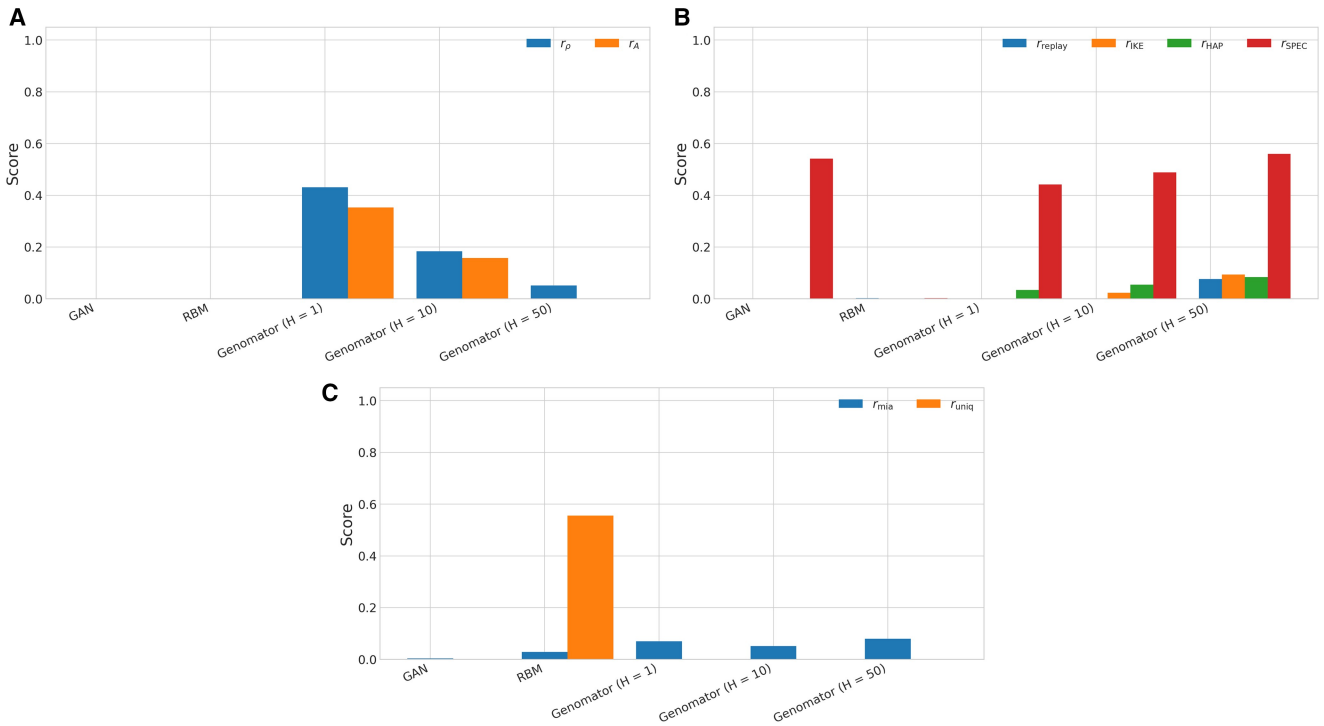


Figure 2 PRISM-G submetrics for Chromosome 1 dataset. Each barplot shows (A) Proximity leakage components. (B) Kinship replay components. (C) Trait leakage components.

Downloaded from academic.oup.com/bioinformatics/article/42/Supplement_2/btag377/8767287 by INACTIVE user on 29 September 2026

intermediate kinship signals ($0.441 < \text{KRI} < 0.560$), and low trait-linked leakage ($0.052 < \text{TLI} < 0.079$).

Results for the Chromosome 15 dataset showed similar qualitative patterns (Supplementary Fig. S2): RBM exhibited the strongest trait-linked leakage, Genomator configurations were mainly driven by proximity and kinship signals varying with the H constraint, and GAN showed comparatively balanced component scores. Full estimates are provided in Supplementary Tables 7–9.

PRISM-G score and privacy ranking

We first estimated PRISM-G scores using the reference anchors with default hyperparameters and performed a sensitivity analysis through grid search to assess the effect of calibration parameters on the estimated privacy risk (Supplementary Data). Under the default configuration, PRISM-G scores for the Chromosome 1 dataset yielded the following ranking from safer to riskier: **GAN > Genomator ($H = 50$) > Genomator ($H = 10$) > Genomator ($H = 1$) > RBM**. Across the sensitivity analysis, ridge regularization primarily affected the dispersion of scores, particularly at higher marker densities, but did not alter the relative ordering of generators. Overall, rankings and qualitative conclusions remained consistent across the explored parameter grid, indicating that PRISM-G estimates are robust to plausible variations in calibration and regularization parameters (Supplementary Fig. S3). Complete estimates from the sensitivity analysis and ranking results are reported in Supplementary Tables 10–17.

Using the hyperparameter configuration that maximized the average Kendall rank correlation across bootstrap replicates, we obtained the final PRISM-G estimates summarized in Figure 4 for the Chromosome 1 dataset. Scores ranged from 1.204 (GAN) to 42.034 (RBM), with Genomator configurations showing intermediate values that decreased as the Hamming distance constraint was relaxed (27.535 for $H = 1$ to 9.628 for $H = 50$). Stratifying the scores by quartiles of the PRISM-G scale, all datasets remained within the privacy-preserving range ($\text{PRISM} - \text{G} < 50$). Rankings remained stable across the parameter grid ($\tau = 0.985$, $\text{P} - \text{value} < 0.018$), consistently identifying GAN as the safest generator and RBM as the most privacy-leaky method. Similar patterns were observed for the Chromosome 15 dataset, although overall risk scores were higher, shifting RBM closer to the leaky reference (Supplementary Fig. S4).

Privacy-utility trade off

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 the procedure described in the Experimental Setup section. Figure 5 shows the obtained performance metrics and privacy-utility frontier for the Chromosome 1 dataset. Across generators, synthetic datasets preserved population structure and achieved high utility for ancestry prediction (Fig. 5A and Supplementary Figs. 5-6).

Pairing the PRISM-G scores with the average performance metrics as a utility measure yields the privacy–utility frontier shown in Fig. 5B. 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 risk. In contrast, RBM is dominated, exhibiting higher privacy risk without a corresponding utility advantage. Similar privacy–utility patterns were observed for the Chromosome 15 dataset (Supplementary Figs. 7-8).

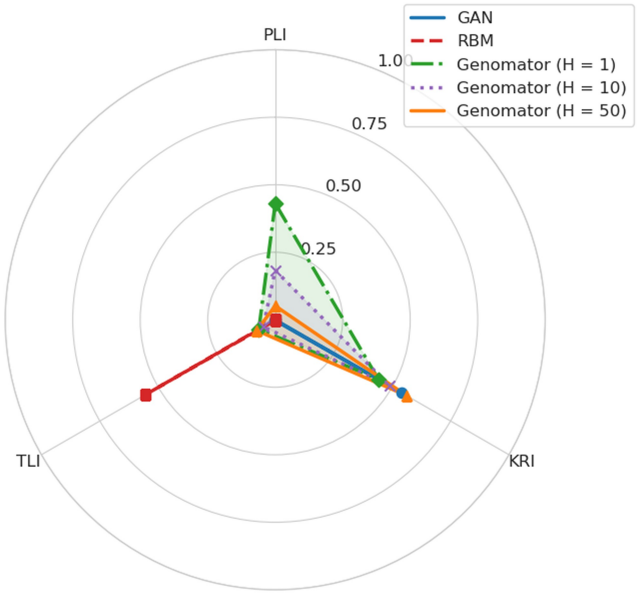


Figure 3 PRISM-G components (PLI, KRI, and TLI) for the Chromosome 1 dataset.

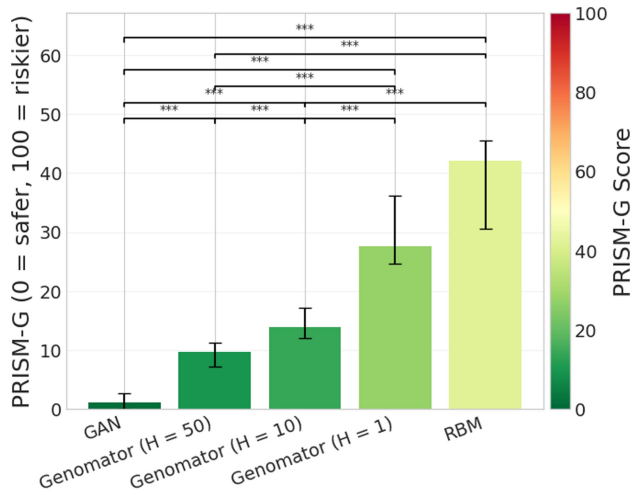


Figure 4 Mean PRISM-G scores for the Chromosome 1 dataset and pairwise comparisons obtained using grid search evaluation.

Discussion

Current evaluations of synthetic genomic data often rely on similarity-based diagnostics to argue that generated genomes differ sufficiently from real samples (Giuffrè and Shung 2023,

Downloaded from academic.oup.com/bioinformatics/article/42/Supplement_2/btag377/8767287 by INACTIVE user on 29 September 2026

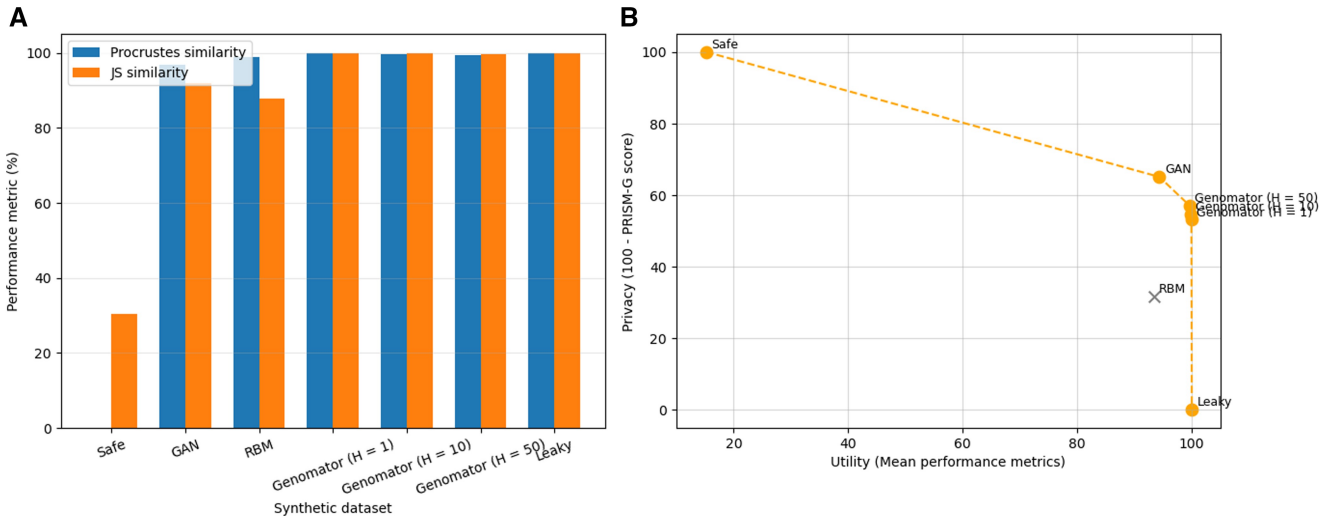


Figure 5 Performance metrics and privacy–utility trade-off for the Chromosome 1 dataset: (A) population structure preservation and ancestry inference utility; (B) privacy–utility frontier based on PRISM-G scores.

Wharrie *et al.* 2023, Kenneweg *et al.* 2025, Royo *et al.* 2025). However, similarity alone does not capture risks such as singling-out, relatedness replay, or rare-variant leakage (Kaabachi *et al.* 2025). To address this gap, we introduced PRISM-G, which evaluates privacy risk across proximity, kinship, and trait-linked signals. Applying PRISM-G across multiple generators and SNP panels from the 1KGP dataset revealed distinct privacy profiles: GAN showed the lowest overall risk, RBM exhibited elevated rare-variant leakage, and Genomator displayed parameter-dependent trade-offs between proximity and relatedness signals. Overall, these findings highlight that privacy exposure arises through multiple mechanisms rather than a single similarity signal, supporting the need for domain-aware privacy evaluation frameworks (Pilgram *et al.* 2025, Zamzmi *et al.* 2025).

Beyond component-level diagnostics, PRISM-G also enables privacy–utility evaluation. Although all generators achieved similarly high utility for ancestry prediction, their privacy risks differed substantially. Genomator maintained high utility while reducing privacy risk as the Hamming distance constraint increased, whereas RBM showed higher privacy exposure without improved analytical performance. These results highlight that generators with similar utility may differ markedly in privacy risk, supporting the use of PRISM-G for model selection and parameter tuning.

Three practical insights emerge from our analysis. First, component-level diagnostics are essential, as similar aggregate scores may arise from different leakage mechanisms. Second, calibration against reference generators provides an interpretable privacy scale that facilitates comparison across models. Third, our results reinforce that privacy risk depends on dataset characteristics and generator settings, indicating that no single model is uniformly safest across contexts.

PRISM-G has several key limitations that constrain interpretation of its results. First, it is an empirical framework without formal privacy guarantees, and its outputs depend on the dataset, selected features, and calibration anchors; thus, **a low score reflects lower relative risk under the chosen setup rather than absence of privacy risk or evidence of non-memorization.** Second, it does not incorporate a demographic null model, so signals such as proximity, haplotype sharing, or IBD may reflect population structure (e.g. admixture, bottlenecks, or shared ancestry) rather than memorization. Third, its aggregation assumes independent leakage pathways, while these signals may be correlated; more flexible approaches such as regression, hierarchical, or copula-based models could better capture such dependencies. Finally, evaluations are limited to subsets of the 1KGP panel, limiting robustness assessment across populations; broader analyses across diverse cohorts and subgroups are needed to characterize generalizability and estimate null and expected metric distributions.

Future work should address these gaps by incorporating population-specific forward or coalescent simulations to better model expected metric distributions under realistic demographic histories, and by extending the framework with additional genomic features, alternative embeddings, richer trait-linked representations (including combinations of common variants), and expanded downstream applications such as genome-wide association studies and polygenic risk score modeling.

Conclusions

Overall, PRISM-G provides a compact and interpretable approach for evaluating privacy risk in synthetic genomic datasets. By combining component-level diagnostics, calibrated risk scores, and privacy–utility analysis, the framework enables systematic comparison of synthetic genome generators towards informed decisions about the safe use and release of synthetic genomic data.

Author contributions

Alejandro Correa Rojo (Conceptualization [Lead], Methodology [Lead], Investigation [Lead], Writing—original draft [Lead], Writing—review & editing [Equal]), Yves Moreau (Supervision [Equal], Writing— review & editing [Equal]), and Gökhan Ertaylan (Supervision [Equal], Writing—review & editing [Equal])

Downloaded from academic.oup.com/bioinformatics/article/42/Supplement_2/btag377/8767287 by INACTIVE user on 29 September 2026

Supplementary material

Supplementary material is available at *Bioinformatics* online.

Conflict of interests

The authors declare that they have no competing interests.

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) and Beyond the Genome: Ethical Aspects of Large Cohort Studies Fonds voor Wetenschappelijk Onderzoek - Junior Project (project ID: FN 701000004).

Data availability

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

References

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

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

Chen Y-M, Hsiao T-H, Lin C-H *et al.* Unlocking precision medicine: clinical applications of integrating health records, genetics, and immunology through artificial intelligence. *J Biomed Sci* 2025;**32**:16. https://doi.org/10.1186/s12929-024-01110-w

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

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

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

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

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

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

Gymrek M, McGuire AL, Golan D *et al.* Identifying personal genomes by surname inference. *Science* 2013;**339**:321–4. https://doi.org/10.1126/science.1229566

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

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**:60. https://doi.org/10.1038/s41746-024-01359-3

Kenneweg P, Dandinasivara R, Luo X *et al.* Generating synthetic genotypes using diffusion models. *Bioinformatics* 2025;**41**:i484–i492. https://doi.org/10.1093/bioinformatics/btaf209

Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* 2009;**5**:e1000384. https://doi.org/10.1371/journal.pgen.1000384

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

Mendes JM, Barbar A, Refaie M *et al.* 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

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

Oprisanu B *et al.* On utility and privacy in synthetic genomic data. arXiv, arXiv:2102.03314 [q-bio], 2022, preprint: not peer reviewed.

Patterson N, Price AL, Reich D *et al.* Population structure and eigenanalysis. *PLoS Genet* 2006;**2**:e190. https://doi.org/10.1371/journal.pgen.0020190

Pilgram L, Dankar FK, Drechsler J *et al.* A consensus privacy metrics framework for synthetic data. *Patterns (NY)* 2025;**6**:101320. https://doi.org/10.1016/j.patter.2025.101320

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**:iii25–iii31. https://doi.org/10.1093/eurpub/ckaf070

Shabani M, Marelli L. Re-identifiability of genomic data and the GDPR. EMBO Rep 2019;**20.**https://doi.org/10.15252/embr.201948316

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**:141–7. https://doi.org/10.1038/s41576-024-00776-0

1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* 2015;**526**:68–74. https://doi.org/10.1038/nature15393

Thomas M, Mackes N, Preuss-Dodhy A *et al.* Assessing privacy vulnerabilities in genetic data sets: scoping review. *JMIR Bioinform Biotechnol* 2024;**5**:e54332. https://doi.org/10.2196/54332

VanRaden PM. Efficient methods to compute genomic predictions. *J Dairy Sci* 2008;**91**:4414–23. https://doi.org/10.3168/jds.2007-0980

Venturini P, Faria PL, Cordeiro JV *et al.* 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

Downloaded from academic.oup.com/bioinformatics/article/42/Supplement_2/btag377/8767287 by INACTIVE user on 29 September 2026

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**:btad535. https://doi.org/10.1093/bioinformatics/btad535

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

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**:e1011584. https://doi.org/10.1371/journal.pcbi.1011584

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

Downloaded from academic.oup.com/bioinformatics/article/42/Supplement_2/btag377/8767287 by INACTIVE user on 29 September 2026

© The Author(s) 2026. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
Bioinformatics, 2026, 42, 1–9
https://doi.org/10.1093/bioinformatics/btag377
25th European Conference on Computational Biology supplement