Private information leakage from polygenic risk scores

Kirill Nikitin and Gamze Gursoy

Genome Res. published online August 17, 2026
Access the most recent version at doi:10.1101/gr.282202.126

Private information leakage from polygenic risk scores

Kirill Nikitin[1,2] and Gamze Gürsoy[1,2,3,4]

[1]Department of Biomedical Informatics, Columbia University, New York, NY 10032, USA; [2]New York Genome Center, New York, NY 10013, USA; [3]Department of Computer Science, Columbia University, New York, NY 10032, USA; [4]**Current Address:** Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, CB3 0WA, UK

Polygenic Risk Scores (PRSs) estimate the likelihood of individuals to develop complex diseases based on their genetic variations. While their use in clinical practice and direct-to-consumer genetic testing is growing, the privacy implications of public PRS sharing are often underestimated. In this work, we demonstrate that PRSs can be exploited to recover genotypes and to de-anonymize individuals. We describe how to reconstruct a portion of an individual's genome from a single PRS value by using dynamic programming and population-based likelihood estimation, which we experimentally demonstrate on PRS panels of up to 50 variants. We highlight the risks of combining multiple, even larger-panel PRSs to improve genotype-recovery accuracy, which can enable the re-identification of individuals or their relatives in genomic databases or the prediction of additional health risks, not originally associated with the disclosed PRSs. We then develop an analytical framework to assess the privacy risk of releasing individual PRS values and provide a potential solution for sharing PRS models without decreasing their utility.

A polygenic risk score (PRS) is a predictive measure that quantifies an individual's genetic susceptibility to complex traits, based on the cumulative effect of their inherited genetic variants. PRSs have demonstrated robust predictive performance across a wide range of complex diseases (Khera et al. 2018; Allegrini et al. 2019), including successful applications in clinical cohorts (Hsu et al. 2015; Mavaddat et al. 2019; Seibert et al. 2018; Desikan et al. 2017; Wray et al. 2018; Musliner et al. 2019). Their utility for risk stratification has spurred discussions about integrating PRS-based assessments into routine healthcare (Mavaddat et al. 2019; Khera et al. 2017; Torkamani et al. 2018). Leading direct-to-consumer genetic testing companies, such as 23andMe (23andMe 2024) and Nebula Genomics (Nebula Genomics 2024), already employ PRSs to provide consumers with personalized trait risk estimates. The use of PRSs in therapeutic decision-making, disease screening, and personalized life planning is anticipated to significantly expand in the coming years (Torkamani et al. 2018).

While ethical and policy debates on the commercial use of PRSs are beginning to take shape (Meyer et al. 2024; Yanes et al. 2024), the privacy implications of sharing individual scores remain largely unexplored. PRSs can be viewed as summary-level data because a single value is assumed to lack sufficient granularity to disclose any additional information beyond the assessed trait's risk. Some genetic association studies publicly release individual PRSs (Ibanez et al. 2017; Kloeve-Mogensen et al. 2021; Thompson et al. 2022), and customers of direct-to-consumer genetic testing companies share their scores on online forums to seek health advice (ApoE4.Info 2023; Reddit 2021b,a). PRSs are, however, derived from genetic data, which have long been regarded as highly sensitive (Lin et al. 2004), due to their potential to re-identify individuals (Jacobs et al. 2009; Pakstis et al. 2010; Gymrek et al. 2013; Erlich et al. 2018) or to reveal their health status (Nyholt et al. 2009; Wang et al. 2009). As the use of PRSs becomes integrated into clinical practice and research, it is essential to understand the privacy implications associated with sharing this information.

In this study, we present the first investigation into the potential leakage of sensitive genetic information through PRSs and evaluate scenarios in which this poses privacy risks. By framing genotype recovery from PRS values as a subset-sum problem, we have developed an efficient dynamic-programming algorithm that uses basic population genetic statistics to accurately infer the underlying genotypes from PRS values. We find that the risk of disclosing patient-level PRS values depends on the parameters of PRS models, specifically the number of loci and the precision of the effect weights, and that models with up to 80 variants are vulnerable to genotype recovery in practice. While PRS models may incorporate thousands of variants, models with fewer than a hundred variants remain common (Gibson et al. 2024; Xu et al. 2024; Marston et al. 2025; Saklatvala et al. 2026) due to their direct biological interpretability and lower cost of clinical implementation.

We first analyze 4,723 PRS models published in the PGS Catalog (Lambert et al. 2021) and find that at least 447 of them would be vulnerable if patient-level PRSs were exposed. We then take a subset of the vulnerable models, each with up to 50 variants, and experimentally demonstrate that, given a panel of PRSs, e.g., from a direct-to-consumer genetic-testing report, an attacker can reconstruct the associated genotypes of an individual with 95% accuracy. Notably, individuals of non-European ancestry appear to be at a greater risk of privacy leakage due to biases in existing genome-wide association studies (GWAS). The recovered genotypes are then sufficient for the re-identification of the individuals or their genetic relatives in genealogy databases with nearly 100% accuracy. We also demonstrate that a typical published PRS based on just 27 loci is sufficient to uniquely identify 95% of individuals in a large cohort of 487K participants. Finally, we develop an analytical framework to assess the privacy risks associated with sharing PRS values and models, and we propose a simple yet effective solution that preserves individual privacy while maintaining the utility of shared information.

We analyze three distinct privacy threat models, each with different adversarial assumptions. First, we study genotype recovery from published PRSs under the assumption that model parameters are publicly available. Second, we examine genealogy-based re-identification that uses recovered genotypes and public genealogy databases. Finally, we evaluate PRS-based linkage risk under an explicit access model where an adversary can compute PRSs for individuals in an anonymized genotype-phenotype database.

Results

PRS-based genetic privacy attacks

A polygenic risk score (PRS) is a numerical value that quantifies an individual's genetic predisposition to a particular disease or trait. It is calculated by aggregating the effects of multiple genetic variants. A PRS is by definition a summary statistic derived from a patient's genetic data, but it is generally considered not to contain any directly identifiable private information about the patient. In this study, we explored whether private genetic variants could be recovered from PRSs and whether these variants could then be used to infer additional sensitive information about patients.

Our attack framework (Fig. 1) is based on two key scenarios: (1) If a research study publicly releases anonymized individual PRSs (Ibanez et al. 2017, Fig. 1)(Kloeve-Mogensen et al. 2021, Supp. Table 5), a genetic-testing facility sells a de-identified PRS dataset (Tanner 2017), or an anonymous user discloses their PRS online to seek health advice (Reddit 2021a,b; ApoE4.Info 2023), it can be possible to infer the associated genotypes of these anonymized patients and to cross-reference genealogy websites to re-identify the patient or their relatives. Note that this is achieved by *querying* genealogy websites, as they typically restrict users from downloading participants' genomic data. (2) If a known patient publicly shares their PRS (e.g., through social media or online forums, such as WebMD), it can be possible to infer their associated genotypes and to search anonymized genotype-phenotype databases with restricted access (e.g., genomic beacons (Fiume et al. 2019) or any database that permits only query-based access) to uncover more sensitive phenotypic information. In both scenarios, further analysis could reveal the predisposition to diseases that the original PRS did not include.

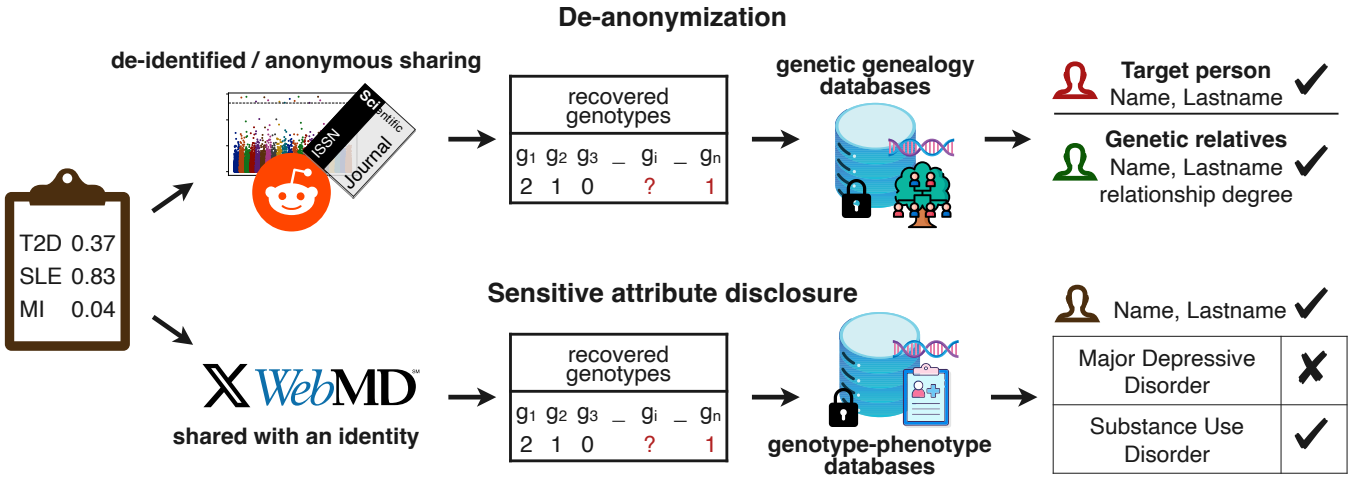

Fig. 1. Overview of privacy risks with sharing PRSs. This illustration depicts the potential privacy risks linked to predicting SNP genotypes from PRS values. For publicly released PRS values of anonymous individuals or from anonymous GWAS participants, the predicted genotypes can be used to de-anonymize them or their genetic relatives by cross-referencing genetic genealogy databases. In cases where PRS values are released for known individuals, the predicted genotypes can be used to infer sensitive health information by searching genotype-phenotype databases.

If a profit-driven entity, such as an insurance provider, gains access to the PRS values and genetic ancestry of a target individual, as well as the corresponding effect weights and SNP loci, both of which can be obtained from resources such as the Polygenic Score (PGS) Catalog (Lambert et al. 2021), it could reconstruct the genotypes for the SNPs included in the PRS panel. These reconstructed genotypes can then be used to: (a) identify the target individual or their genetic relatives in genealogy databases, (b) infer other potential diseases that the target might have by cross-referencing query-only genotype-phenotype databases, or predict predisposition for other diseases of the individual. In all cases, sensitive information about the target is exposed, either revealing their identity or disclosing private health details.

Our goal is to quantify the extent to which an individual's PRS reveals sensitive information, to identify the conditions that amplify or mitigate this exposure, and to develop strategies to reduce associated privacy risks. To achieve this, we developed methods that leveraged publicly available data—such as lists of PRSs, genomic loci of SNPs included in the PRS models, and their corresponding effect weights—to infer an individual's associated SNP genotypes. A PRS for a disease is the linear combination of the genotypes of the associated SNPs, defined as $PRS = \sum \beta_j \cdot g_j$, where β_j is the effect weight of SNP j and g_j is the genotype of SNP j (0 for homozygous reference alleles, 1 for the heterozygous alternative allele, and 2 for homozygous alternative alleles). Inferring the genotypes from a PRS value and effect weights can be framed as a variation of the subset-sum problem (Karp 1972). The subset-sum problem is a computational problem where, given a set of numbers, the goal is to find a subset whose sum equals a target value. It is NP-hard (Garey and Johnson 1979), meaning that it can be computationally difficult to solve, especially as the set size grows. Here, we reduce the genotype recovery task to an instance of the subset-sum problem by defining the effect weights as the number set and the PRS value as the target sum. The number of times the solution contains each effect weight (0, 1, or 2) determines the genotype for each corresponding SNP.

Our framework for inferring a target individual's SNP genotypes based on a PRS value and the associated effect weights consists of several steps. First, it assesses the solvability of the problem by using the concept of *density* (Lagarias and Odlyzko 1985). In subset-sum problems, density represents the ratio between the size of the set (i.e., the number of weights in our case) and the length of the bit representation of the largest weight. Higher density indicates that there are many weights relative to the bit length of the largest weight,

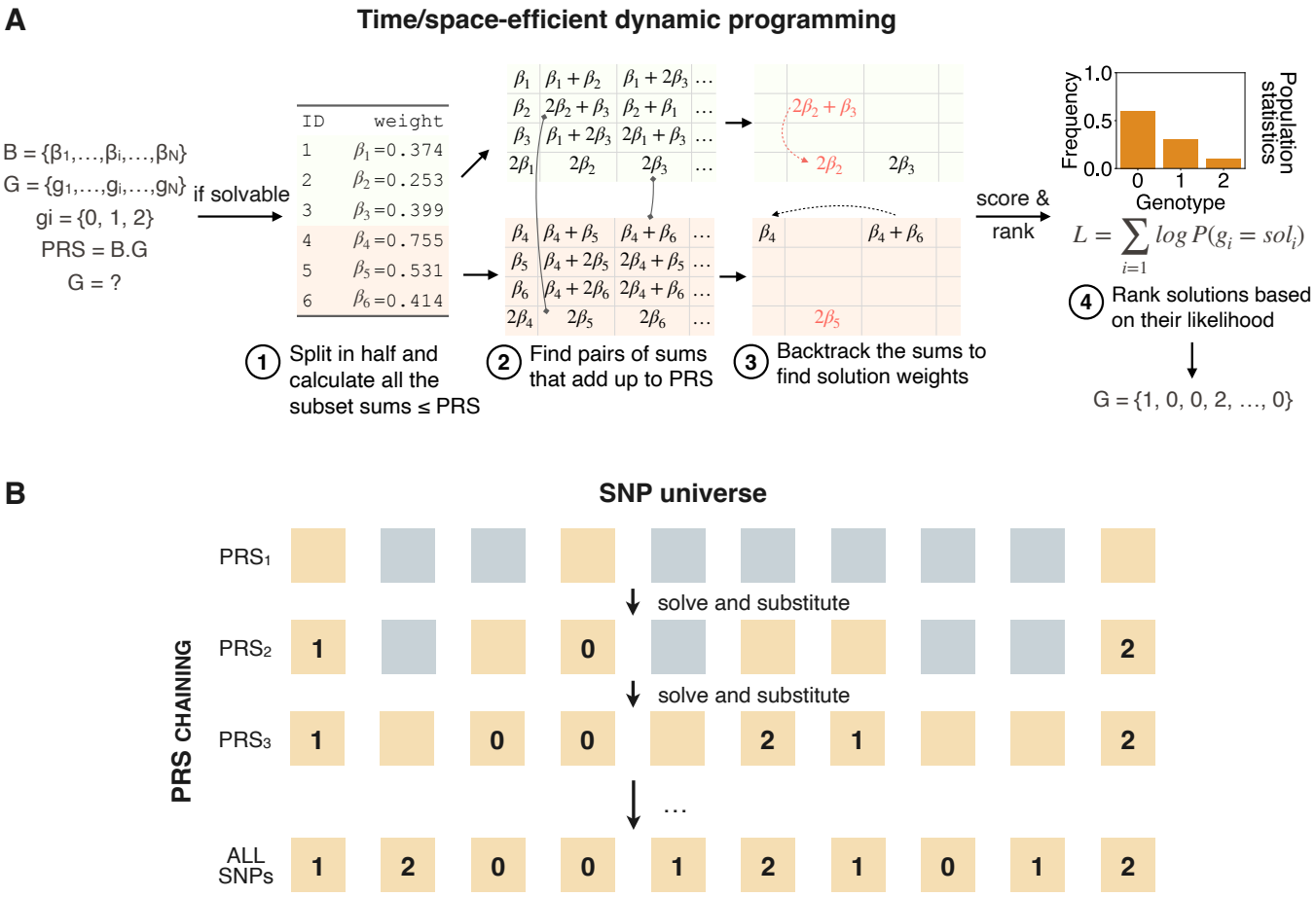

Fig. 2. Framework for predicting genotypes from PRS values and PRS chaining. (A) Illustration of the genotype prediction process via dynamic programming. Given a set of effect weights and a target PRS, the goal is to find the unknown genotypes. We first determine the solvability of a PRS through density calculation. If the PRS is solvable, the framework uses a dynamic programming algorithm to find all possible solutions by using a meet-in-the-middle approach. The meet-in-the-middle approach divides the effect weights into two parts, creating separate tables of partial sums for each. It then identifies pairs of entries from these tables that sum to the target PRS value. Partial sums are added only if the target remains achievable. The search is further refined in two stages: first, rounded-down weights identify solution candidates, and then, precise matches to the target are selected. These strategies significantly reduce the memory and computational costs, which facilitates genotype prediction from PRSs with a larger number of variants. We then score and rank potential genotype solutions based on population allele frequencies. **(B)** Schematic of our PRS-chaining technique. Starting with a PRS involving a small set of SNPs, the predicted genotypes for overlapping SNPs are substituted into progressively larger PRSs. This iterative process narrows the solution space and thus increases the efficiency of genotype prediction as more SNPs are incorporated across multiple PRSs.

which leads to a tighter distribution of possible subset sums. This dense spacing increases the number of combinations that could potentially match the target sum, thus making the problem more challenging to solve (Austrin et al. 2016). We adapted this concept to genetic data by applying a density threshold to determine whether a PRS is solvable—that is, whether we can expect to accurately infer the genotypes for all the associated SNPs of a target individual, given the effect weights and the PRS value (see Methods). Naturally, a PRS with fewer associated SNPs or more precise weights is easier to solve, as the computational problem becomes sparser and more tractable.

Once the PRS is determined to be solvable, we proceed to infer the associated SNP genotypes. To achieve this, we developed a dynamic programming algorithm (Bellman 1957), enhanced by the meet-in-the-middle approach and

further optimized by using additional techniques (Horowitz and Sahni 1974) (Fig. 2A). This algorithm efficiently identifies all valid solutions, i.e., all possible genotype sets that result in the target PRS. In order to assess the plausibility of each potential solution, we calculate a log-likelihood score by using allele frequencies from the target individual's population. Specifically, we compute the sum of the log-probabilities for the identified genotypes in each solution. The core idea is that a solution that closely aligns with the population average is more likely to be correct. The solution with the highest total sum, which corresponds to the highest likelihood, is then selected as the most probable genotype configuration (Fig. 2A). We then incorporate continuous log-likelihood estimation directly into the dynamic-programming algorithm, which enables us to handle PRSs with more SNPs. Additionally, we implement two techniques to further increase the solving accuracy and the size of solvable PRSs, which we call PRS chaining and self-repair. In PRS chaining, we first solve the PRS with the smallest set of SNPs. For each subsequent PRS that includes more SNPs, we retain the genotypes of overlapping SNPs from the previous solution. This approach gradually reduces the possible solution space as the number of SNPs increases, thereby making the computations more efficient (Fig. 2B). When incorporating solutions from a previous PRS, if the solution for the current PRS fails, it suggests that the integrated genotypes might be incorrect. This enables us to revisit and correct these genotypes, a process that we refer to as *self-repair*. For details, see Methods.

Patient genotypes can be recovered from publicly available PRSs

The solvability of the genotype recovery problem is determined by the parameters of the corresponding PRS model, specifically the number of associated SNPs and the precision (i.e., the number of decimal digits) of the effect weights. By following the literature on the subset-sum problem and via experimental evaluation, we determined the range of parameters where genotype recovery was feasible (see Methods). We then analyzed all PRS models published in the PGS Catalog (Lambert et al. 2021) to assess the proportion that could be resolved if individual PRSs were available. We found that at least 447 out of 4,723 published PRS models were vulnerable to genotype recovery.

Correctly inferring genotypes from a single PRS might already suffice for de-identifying an individual, as prior research demonstrated that as few as 20 SNP genotypes could enable re-identification (Lin et al. 2004; Pakstis et al. 2010). In practice, an attacker can obtain a list of PRS values for a single individual, e.g., an insurance company might receive a patient's genetic-testing report with over 40 risk scores (23andMe 2024). Combining prediction results from multiple PRSs, even those with more SNPs than theoretically solvable, enables an attacker to achieve substantial coverage of an individual's genome.

We evaluated the attacker's genotype recovery ability by using whole-genome sequencing data from the 1000 Genomes Project (2,535 samples) (The 1000 Genomes Project Consortium 2015b) and a PRS set containing scores of up to 50 SNPs each from the PGS Catalog for various diseases. After filtering the PRSs based on our solvability threshold and performing quality controls (see Methods), we selected a set of 298 PRSs, encompassing 4,821 SNPs, of which 2,654 were unique (Supplemental Table S1). We then predicted the genotypes of these SNPs for each individual given their PRS values. We compared the results against two baselines: the deterministic predictor that assigned the major genotype of the individual's ancestry at each SNP, and the stochastic predictor that sampled genotypes according to ancestry-specific probabilities.

Our method achieved a median genotype prediction accuracy of 94.6% (Fig. 3A), significantly outperforming both baseline approaches. On average, we correctly predicted 2,450 genotypes per individual. Notably, we observed higher prediction accuracy for individuals of African (AFR) and East Asian (EAS) genetic ancestries compared to those of European (EUR) ancestry. We hypothesized that this increased accuracy resulted from a higher proportion of loci where the effect allele frequencies (EAFs) were close to 0 or 1 in the AFR and EAS populations.

To test this hypothesis, we analyzed genotype recovery as a function of population EAF. We confirmed that recovery was

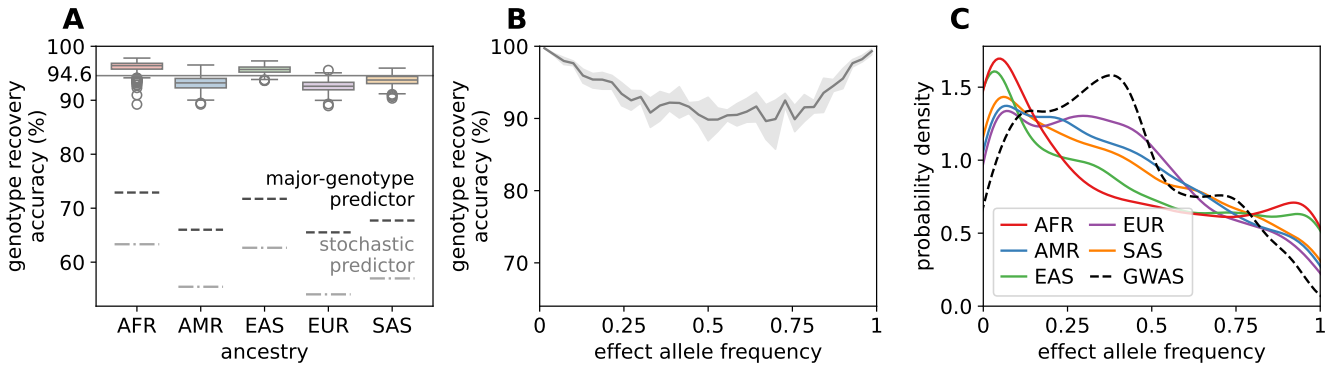

Fig. 3. Genotype recovery accuracy. **(A)** Genotype recovery accuracy across different populations: African (AFR), Admixed American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS). The baselines represent the recovery accuracy of predicting the major genotype or stochastically sampling a genotype for each SNP for every individual, according to the population allele frequencies, without access to the PRS values. **(B)** The relationship between genotype recovery accuracy and effect allele frequency. It is easier to predict genotypes for SNPs when either allele is rare. **(C)** The distribution of effect allele frequencies in different populations compared to those in GWASs from which the PRSs are derived. Since individuals of EUR genetic ancestry dominate most GWAS datasets, the effect allele frequency distribution in GWASs closely mirrors that of the EUR population.

more challenging for loci with EAFs near 0.5 (Fig. 3B), where higher heterozygosity complicated genotype inference. In contrast, EAFs near 0 or 1 indicated reduced genetic variation and higher homozygosity, which increased the likelihood of correctly predicting that an individual carries the predominant allele.

We investigated why individuals of AFR and EAS ancestry carried more effect alleles at the loci where the allele frequencies in their populations were close to 0 or 1. We found that this was primarily due to the over-representation of PRSs derived from GWASs that had been conducted predominantly on cohorts of EUR ancestry (Fig. 3C), which aligned with prior findings on clinical study disparities (Martin et al. 2017). The allele frequencies in the EUR population closely mirrored the effect allele frequencies in the GWASs. This bias led to significant discrepancies between the effect allele frequencies reported in the GWASs and the actual allele frequencies in the AFR and EAS populations. As a result, individuals from these populations were more susceptible to genotype recovery attacks because their genotypes at these loci were more predictable.

We then asked whether the high genotype recovery accuracy depended on the assumption that the attacker knew the target individual's genetic ancestry. To test this, we selected ten individuals from each ancestry group. We then performed genotype recovery with our dynamic-programming approach and scored solutions with likelihood estimates based on universal, ancestry-agnostic allele frequency. The median accuracy across all fifty individuals was 92.7%, a slight decrease from the accuracy obtained with ancestry-specific data (Supplemental Figure S2). As expected, individuals of the AFR and EAS ancestries showed the largest drop in accuracy, with a decrease of 3%. These results indicated that, while precise population statistics enhanced genotype recovery accuracy, our method remained robust even when the attacker lacked knowledge of the target individual's specific ancestry.

Patients and their relatives can be identified in genealogy databases by using PRSs

If PRSs are publicly shared for anonymized study participants (Thompson et al. 2022) or posted online by anonymous users seeking health advice (Reddit 2021a,b; ApoE4.Info 2023), there is a potential risk that an attacker could re-identify the individuals. To test this risk, we designed an attack strategy that involved querying a genetic genealogy database, such as GEDMatch (Ram and Roberts 2019), with the genotypes inferred from published PRSs. This approach could potentially reveal the identity of the tar-

get individual or their close relatives. It is important to note that genealogy databases do not provide direct access to individuals' genomes; instead, they allow users to query the database to identify genetic relatives.

Although genealogy services do not publicly disclose the specific algorithms that they use for identifying relatives, prior studies (Ney et al. 2020; Edge and Coop 2020) suggest that these services search for uninterrupted genomic segments with no mismatching homozygous sites between pairs of individuals. To assess whether the predicted genotypes from our genotype recovery process were sufficient for de-anonymization, we conducted a relative search process by using the 1000 Genomes dataset as the genealogy database. This database contains genomes from 2,467 unrelated individuals, and 50 and 18 individuals with first- and second-degree relationships, respectively. We used the KING-robust algorithm (Manichaikul et al. 2010) as the relatedness inference method. It uses a large set of autosomal SNPs to calculate kinship coefficients, which help identify relationships, such as first- or second-degree relatives. The exact number of SNPs that it uses can vary depending on the dataset, but having a large number of SNPs enhances the accuracy of kinship coefficient estimates.

Despite our genotype predictions being based solely on SNPs inferred from the PRSs, we demonstrated that they were sufficient to achieve 100% precision and recall in identifying individuals, i.e., our test correctly identified all 2,535 individuals as themselves. We also achieved ~90% precision and recall in determining first-degree relatives, and ~85% precision and ~75% recall in determining second-degree relatives. This contrasts with the 0% accuracy of a baseline approach that simply predicted the major genotype for all the associated SNPs (Fig. 4B). While our genotype predictions were not perfectly accurate and covered only about half as many SNPs as were considered sufficient for relatedness analysis, they still provided enough information to distinguish different levels of relatedness. The matching precision was comparable between individuals of all ancestries, thus a higher genotype prediction accuracy did not translate into a higher risk of re-identifiability, as predicted SNPs that were common within a population contributed little to individual distinctiveness.

Kinship coefficients are an absolute metric; therefore, their performance does not directly depend on database size. The distribution of the KING kinship coefficients from our analysis in Supplemental Figure S3A indicates a clear separation between self-identification and unrelated individuals or even first-degree relatives. This mirrors the relationship between relatedness and genotype density observed in the original KING study (Manichaikul et al. 2010). In order to evaluate the re-identification performance in a large-scale setting, we perturbed the genotypes of the individuals in the UKBB dataset according to the error distribution of the genotype recovery and linked them to the full database of approximately 487K samples. The results in Supplemental Figure S3C demonstrate robust linking performance for individuals and monozygotic twins. The kinship coefficients for first- and second-degree relatives were still, on average, significantly above those of unrelated individuals but did not match the KING relatedness cut-offs. Finally, we replicated the analysis using GCTA (a statistical method for heritability estimation (Yang et al. 2011)) on the predicted SNPs of the 1000 Genomes individuals and found that, although GCTA exhibited slightly lower accuracy, it still successfully identified individuals and most of their relatives as related (Supplemental Figure S3B).

Anonymized genotype databases can be de-anonymized by using a single PRS from known patients without the need for genotype prediction

Previous studies have shown that even a few dozen SNPs can uniquely identify an individual (Lin et al. 2004; Pakstis et al. 2010). Building on this, we examined whether a single PRS could distinguish an individual within a large genotype-phenotype database. In this scenario, we assume direct access to an anonymized genotype-phenotype database, such as UK Biobank, which insurance companies might have access to (The Guardian 2023). We tested whether, given an anonymized genetic database, the PRS of a known individual for a particular trait or disease, and the effect weights of the SNPs in the PRS model, it would be possible to calculate

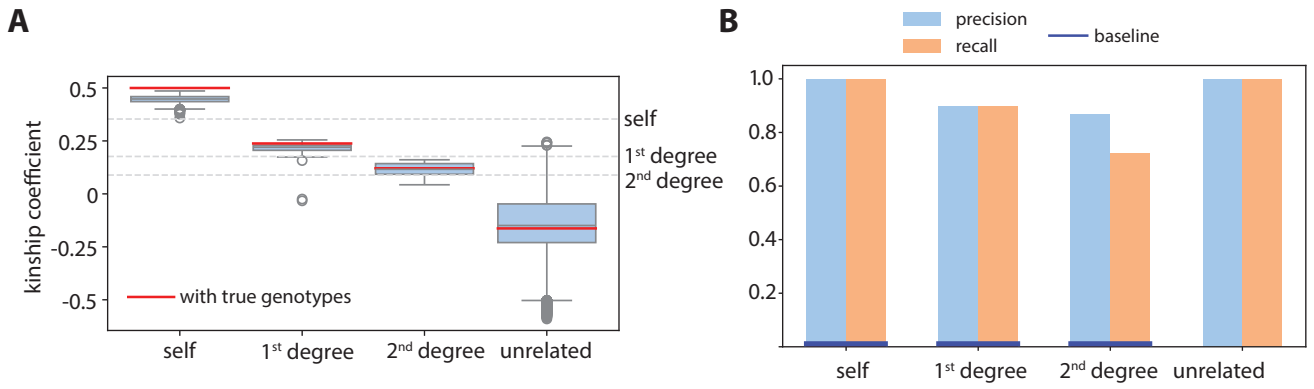

Fig. 4. Comparison of kinship coefficient estimations and relationship classification performance in the 1000 Genomes Database. **(A)** Distribution of kinship coefficients for different relationship categories (self, 1st degree, 2nd degree, and unrelated) calculated with KING-robust using the predicted genotypes. The red lines are calculated using the correct genotypes for the same SNPs. **(B)** Precision and recall metrics for identifying self, 1st degree, 2nd degree relationships using the predicted genotypes. The baseline performance reflects the re-identification approach of predicting the major alleles for the SNP genotypes.

PRS values for all individuals in the database and identify the matching one (Fig. 5A). This approach relies on determining how unique PRS values are across individuals, raising the possibility that an attacker could use a single PRS to link a known individual to an anonymized dataset and to potentially expose sensitive genetic and phenotypic information. The viability of this method depends on whether the PRS for a given trait or disease is sufficiently unique to match a specific individual in the dataset.

We first asked whether PRS values were unique to individuals. For this, we developed a theoretical estimation of the number of uniquely identifiable individuals in a database based on the database's size, PRS parameters (effect weights and their precision, i.e., the number of decimal places), and allele frequency statistics in the population. As the majority of genotypes behave as independent random variables (SNPs can be in most cases considered independent due to the clumping step in PRS modeling (Pain et al. 2024)), their linear combinations, given a sufficient sample size, follow a normal distribution, as described by the central limit theorem. By incorporating the effect weights and allele frequencies from the population, we could estimate both the mean and standard deviation of this distribution (see Methods). We showed that the total number of possible scores for a given PRS model was determined by either the number of SNPs or by the precision of the effect weights involved in that specific PRS, whichever yielded a smaller value. By using this distribution, we calculated the proportion of scores uniquely attributable to individuals ("uniqueness"), as well as the median number of individuals sharing the same score ("the median anonymity-set size"). The anonymity-set size refers to the number of individuals in a group who are indistinguishable from each other based on available data. We compared our analytical estimates with the real scores calculated for the 1000 Genomes and UKBB individuals on the PRSs selected from the PGS Catalog for this study. We found that a single PRS model with 20 SNPs in 1000 Genomes or 27 SNPs in UKBB, on average across all weight precisions, sufficed to uniquely identify 95% of individuals (Fig. 5B). Even when the scores overlapped, the same score often was shared by only a few individuals. For instance, in a PRS model with 7 SNPs (1000 Genomes) or 14 SNPs (UKBB), the median anonymity-set size was two, meaning that half of the scores were shared by at most two individuals (Fig. 5C). See Supplemental Figure S4 for the relationship between the number of possible scores and the number of SNPs observed in the actual data (1000 Genomes and UKBB).

Our analysis slightly overestimates the number of uniquely identifiable individuals, as it does not account for the repetition of effect weights or residual linkage disequilibrium

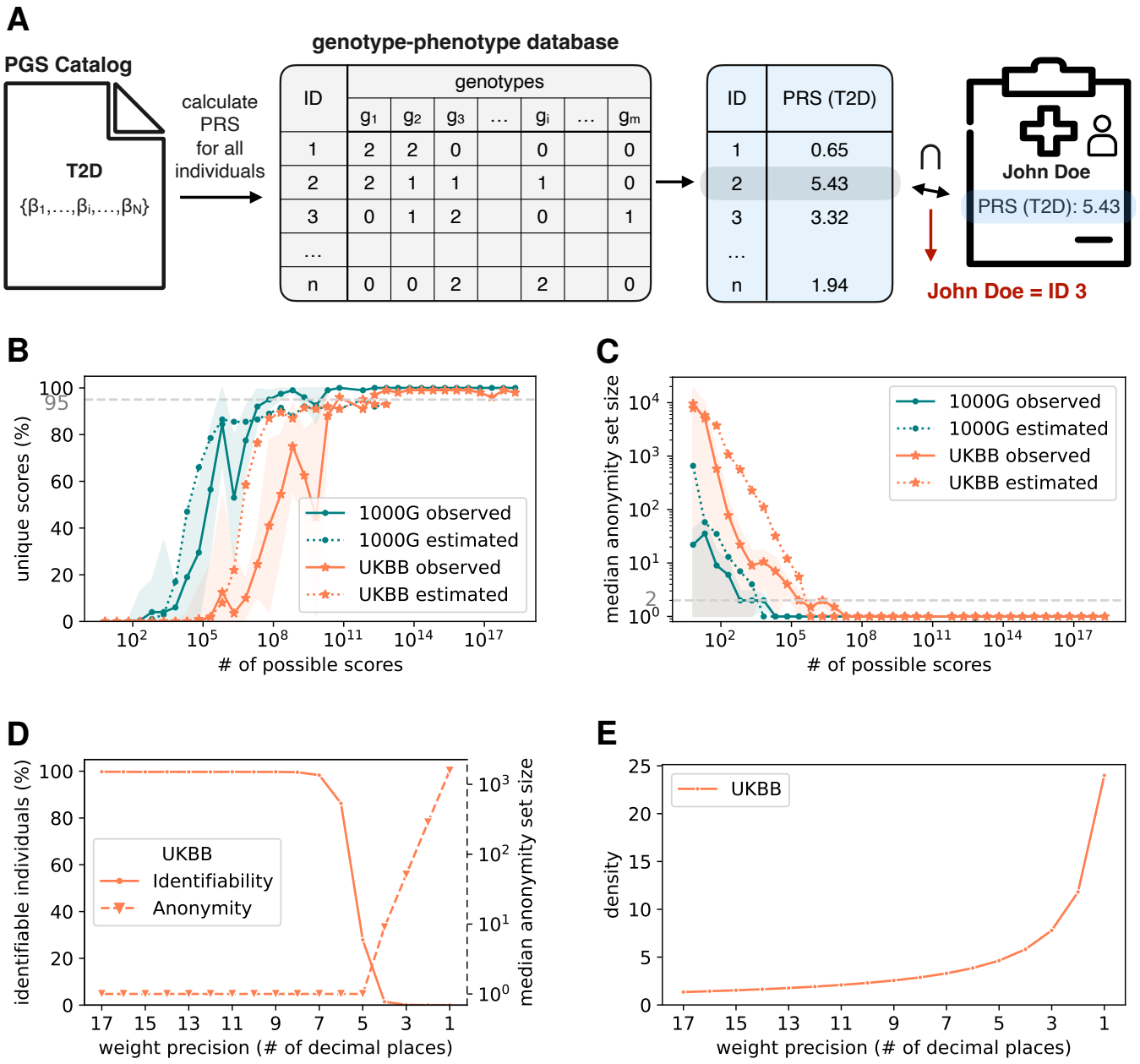

Fig. 5. Uniqueness of PRS values. **(A)** Linking released PRS values to an anonymized genotype-phenotype database by taking advantage of the uniqueness of the PRS values. **(B)** Percentage of unique PRS values as a function of number of possible scores both observed and estimated using the 1000 Genome and UKBB datasets. We found that 95% uniqueness achieved in UKBB with $\sim 10^{11}$ possible scores, which corresponds to a PRS composed of 27 SNPs. **(C)** Median number of individuals who share the same PRS values as a function of number of possible scores. We found that, on average, two individuals share the same PRS value in UKBB with $\sim 10^{6}$ possible scores, which corresponds to a PRS composed of only 14 SNPs. **(D)** The percentage of identifiable individuals and the median number of individuals who share the same PRS value as a function of number of decimal places reported in effect weights. As we round the effect weights, we see a drastic decrease in the number of identifiable individuals and an increase in the median number of individuals who share the same PRS value. This plot is calculated using a PRS model with 48 SNPs for Type-I diabetes (PGS000869 from the PGS Catalog (Aksit et al. 2020)) and UKBB individuals. **(E)** Density of subset-sum problem for the genotype recovery using PRS value as a function of number of decimal places reported in the effect weights. As we round the effect weights, genotype recovery becomes a difficult problem to solve, *i.e.*, the density increases. This plot is calculated using a PRS model with 48 SNPs for Type-I diabetes (PGS000869 from the PGS Catalog (Aksit et al. 2020)) and UKBB individuals.

in PRS models, which leads to deviations from the normal distribution and makes some scores more common than predicted. It also pertains to large mixed-cohort databases where the allele distribution conforms to the general population. A disease-specific database might include individuals with a higher rate of risk alleles and hence PRSs at the tail of the distribution, which would decrease identifiability. However, the analysis accurately captures the overall trend, which enables either an attacker to estimate the potential success of their de-anonymization attack before accessing a dataset or a data custodian to perform risk assessment. Across multiple traits, the probability of two individuals sharing identical PRSs decreases multiplicatively, which makes a combination of several PRSs a robust identifier even in the largest available datasets.

This analysis considers a distinct linkage threat model in which an adversary has access to an anonymized genotype database—or to PRS values derived from such data—and seeks to link individuals across datasets by using PRS uniqueness. This access may arise through unauthorized database access, insider misuse, or policy-compliant research use in which identifiers have been removed but genetic data remain available. This access assumption applies only to the PRS-uniqueness analysis presented here and is not required for the genotype-recovery or genealogy-based scenarios analyzed in previous sections. The purpose of this analysis is to quantify identifiability under these assumptions. While statistical uniqueness alone does not guarantee exploitability in real-world settings, it measures the potential for linkage if such access is available.

Finally, we investigated a potential strategy to enhance privacy. As the uniqueness of a PRS largely depended on the precision of the effect weights, reducing the number of decimal places in effect weights decreased the uniqueness of each PRS and increased anonymity-sets sizes(Fig. 5D). In other words, if PRS values were calculated by using precise weights but the released PRS data were rounded, it became more challenging to match known individuals to anonymous datasets. We further demonstrated this by analyzing PRS solvability and showing that rounding effect weights increased the density of possible solutions (Fig. 5E) and, thus, hindered genotype prediction from the PRS. While rounding effect weights greatly improved privacy, it had almost no impact on the utility of the PRS, as the distributions of PRS values calculated with precise and rounded weights remained closely aligned (Supplemental Figure S5).

Discussion

In this study, we identify critical privacy risks associated with the sharing of PRSs. By framing genotype inference from PRSs as a subset-sum problem, we demonstrate that individual genotypes can be accurately predicted from PRSs, achieving a median accuracy of 94.6% across diverse ancestry groups. This capability to reconstruct genetic information from PRSs raises serious privacy concerns, particularly when PRSs are openly shared in research publications, disclosed by individuals seeking health insights, or sold in a de-identified form by genetic-testing companies.

Our experiments evaluate genotype recovery from PRSs with up to 50 SNPs, with an estimated upper bound of 80 SNPs. These limits could be extended by increasing available computational resources or by using finer-grained population information, e.g., subpopulation-specific allele frequencies or linkage disequilibrium statistics, to identify the correct genotypes among valid solutions. However, even small-scale PRS models are not rare—approximately 12% of the models listed in the PGS Catalog that we analyzed contained fewer than 50 variants.

We further reveal that the risk of inferring genotypes from PRSs is not uniform across populations. Individuals of African and East Asian descent are particularly vulnerable due to differences in allele frequencies between these groups and the European-centric datasets commonly used to construct PRS models. This underscores the need to consider the diverse genetic backgrounds represented in genomic datasets when accessing the privacy risks of PRS models. Additionally, our analysis shows that PRS values can be used to predict health risks beyond the traits originally included in the PRS. This ability to infer supplemental phenotypic information poses challenges for patient consent and data governance frameworks, which often overlook such

secondary uses of data. As PRS-based predictions become more integrated into clinical decision-making, a reassessment of current informed consent policies might be needed to ensure that they address these emerging risks.

To address these concerns, we propose a simple, yet effective mitigation strategy. Reducing the precision of effect weights in publicly released PRS models can decrease the identifiability of individuals while still preserving the overall utility of the scores. Concretely, we propose density as a practical tool for selecting PRS disclosure precision. Data holders can progressively round effect weights, compute the resulting density, and determine whether the model enters a regime where likelihood estimation becomes ineffective in identifying the correct genotype solutions. Our results show that modest rounding substantially reduces recovery accuracy while still preserving overall PRS distributions across different precision levels (Supplemental Figure S5). Density should be interpreted, however, as a conservative indicator of practical risk rather than a formal guarantee. We also note that, while effect-weight rounding substantially reduces privacy risk, high-precision weights are routinely produced and disseminated for reproducibility and reuse. A possible approach for studies is to release two versions of models: one with high-precision weights for reproducibility and another with rounded weights for clinical use. Alternatively, studies could retain high-precision model weights but round individual PRSs whenever they are shared, with the degree of rounding also determined by density.

Finally, integrating privacy-enhancing techniques, such as differential privacy, into the computation and sharing of PRSs might offer a safeguard against re-identification, though care must be taken to balance privacy protection with the preservation of PRS utility. Further research is needed to refine these approaches and establish best practices for maintaining privacy while supporting the use of PRSs in personalized medicine. As PRS evaluations gain wider adoption in healthcare and research, our study emphasizes the need for a deeper understanding of the privacy risks that they pose. It is essential to develop strategies that facilitate the continued use of genetic data for scientific and clinical progress while safeguarding privacy.

Methods

Overview of datasets

In all our experiments, we use publicly available genomic datasets, specifically the phase 3 release of the 1000 Genomes project (2,535 samples) (The 1000 Genomes Project Consortium 2015a) for privacy attacks and the imputed version of the UKBB dataset (487,409 samples) for uniqueness and anonymity estimations (Bycroft et al. 2018). The 1000 Genomes dataset comprises 2,467 unrelated individuals, and 25 and 9 pairs of first- and second-degree relatives, respectively. We use the relatedness information for de-anonymization.

We obtain PRS metadata from The Polygenic Score (PGS) Catalog (Lambert et al. 2021), an open database of published PRS models. The catalog provides metadata for a PRS as a scoring file. We calculate the target PRSs and population allele statistics from the datasets' genotype data, and we use the ancestry and kinship information from the datasets' metadata where appropriate. We do not attempt to infer any unknown attributes or identities of the individuals.

Preprocessing of PRS models from the PGS Catalog

We collected all available PRS models from the PGS Catalog with the number of variants $N < 50$. We filtered them by first removing models that would unfairly simplify our task: PRSs with a single variant $N = 1$ and with large differences in precision across their effect weights. We then removed the PRSs with (1) missing or incorrectly specified loci; (2) effect alleles that did not match either the reference or alternative allele in the 1000 Genomes data; (3) variants on the X or Y chromosomes, or (4) density $d > 2.5$. In total, we ended up with 298 PRSs spanning 4,821 total loci, out of which 2,654 were unique (see Supplemental Table S1 for details).

Genotype prediction as the subset-sum problem

A polygenic risk score is a linear combination of the genotypes $g \in \{0,1,2\}$ of the SNPs associated with a disease and the corresponding effect weights β. The score is typically normalized by the number of SNPs N and the ploidy P. For

diploid cells, $P = 2$:

$$PRS = \frac{\sum_{i=1}^{N} \beta_i g_i}{P \cdot N} \quad (1)$$

The effect weights are real numbers that represent the contribution of each genetic variant (often a single nucleotide polymorphism, or SNP) to the risk of developing a particular disease. Studies release effect weights with varying degrees of precision (the number of decimal places).

The task of finding the genotypes that yield a given PRS value is a variation of the subset-sum problem, in which one attempts to find a subset of a set of numbers that sums to a target value. This problem is NP-hard (nondeterministic polynomial-time hard) (Garey and Johnson 1979) and, hence, can be computationally difficult to solve. However, there are algorithms that can solve this problem efficiently when certain conditions are met regarding the properties of the numbers involved (Bellman 1957; Lagarias and Odlyzko 1985; Coster et al. 1992; Radziszowski and Kreher 1988), and these conditions can be met in the context of PRSs. We hence reduce the genotype-recovery task to an instance of the subset-sum problem by defining the effect weights as the set of numbers and the PRS value as the target sum. The number of times each effect weight (0, 1, or 2) is used in the solution determines the genotype for each SNP.

Solvable PRS

Solving an NP-hard problem even with an efficient algorithm is a computationally intensive process when the problem's size is large. Hence, it is important to be able to estimate the feasibility of genotype recovery before attempting to solve it. The subset-sum problem's difficulty is often measured by using a concept called *density*. This density is defined as the ratio between the number of elements in the set and the bit length (or the number of binary digits) of the largest number in that set (Lagarias and Odlyzko 1985). The idea is that (1) the more numbers a set has, the more possible subsets can be formed, and (2) the shorter the bit length of the numbers, the lower the precision and the fewer possible sums that can be achieved. Naturally, problems with a higher density are more difficult to solve (Austrin et al. 2016) because multiple subsets can yield the same sum.

In contrast to the classic subset-sum problem where the numbers are a set of integers, PRS effect weights are real values and, hence, not all of them can be precisely represented with bits. In addition, genotypes can be one of the three possible values, so we need the ternary bit representation of the largest weight. We adapted the bit precision estimation to the PRS context by calculating the base-three logarithm of the number of decimal places in the largest weight. The number of decimal places must be normalized across all the weights, i.e., if any of the other weights are "longer", we append zeros to the largest weight to match the decimal length. The density is then calculated as the ratio between the total number of effect weights and the largest weight's ternary-bit representation length:

$$d = \frac{N}{\log_3(max_{1 \le i \le N}\, \texttt{decimal}(\beta_i))} \quad (2)$$

Prior work has primarily focused on low-density problems (Lagarias and Odlyzko 1985; Radziszowski and Kreher 1988; Coster et al. 1992; Austrin et al. 2016), i.e., $d < 1$, as most studies considered applications to cryptography, where problems have a single solution. Given our approach of combining dynamic programming with likelihood estimation (see below), we hypothesized that we would be able to solve denser problem instances. To determine a feasible threshold, we selected 268 PRS models of up to 30 SNPs and performed genotype recovery for 50 individuals from the 1000 Genomes dataset for different densities by gradually rounding the effect weights of each model. Our results showed (Supplemental Figure S6) that for density $d > 2.5$, the recovery advantage over predicting the major genotype for each SNP becomes small, which was consistent with the performance of the best-known algorithms (Schroeppel and Shamir 1981; Austrin et al. 2015). We note, however, that density should be used only as a conservative indicator of practical solvability and not as a guarantee of uniqueness or correctness of recovered genotypes.

Resource and density constraints ultimately limit the size of a PRS instance that our algorithm can handle. The computational and memory requirements grow superlinearly with the number of SNPs N and their precision. In order to analyze the requirements in practice, we experimentally measured the CPU time and peak memory usage of genotype recovery across the previously selected 298 PRS models (see Supplemental Figure S7). In this work, we focus on PRS instances with $N < 50$, but we estimate that an adversary with sufficient computational resources could extend this limit to $N < 80$. Most studies choose effect-weight precision with fewer than 15 decimal places, and, hence, the problem becomes too dense to be easily solvable after 80 SNPs.

Our Dynamic Programming Approach

We chose to use dynamic programming to solve genotype recovery posed as a subset-sum problem as it had the best worst-case performance (Bellman 1957). Other efficient algorithms (Lagarias and Odlyzko 1985; Coster et al. 1992; Radziszowski and Kreher 1988) were either based on heuristics, and hence did not guarantee finding the best solution, or were applicable only to low-density problems. We combine dynamic programming with several optimization techniques (Horowitz and Sahni 1974) to reduce the computational costs. We first split the effect weights into two equal-size subsets and build a sums table for each. The process consists of iteratively adding each effect weight (either once or twice to account for one or two effect alleles at the SNP) to the existing weight sums in the table. During each addition, a pointer is recorded to track the last weight added for each sum. This method enables us to construct all the weight sequences that produce each sum in the table. Once the tables are constructed, we use the meet-in-the-middle technique of iterating over all the sums in the first table and checking if the difference between the target and a sum matches any sum in the second table. When we find a match, we backtrack through the tables to find the corresponding weight sequences and translate them into the genotypes, whose linear combination sums to the target PRS.

The meet-in-the-middle technique enables us to reduce the memory complexity from $O(3^N)$ of the direct approach to $O(3^{N/2})$. In order to reduce the costs further, we maintain upper- and lower-bounds on the intermediate sums that still can lead to the target score, given the remaining weights. If a sum is either larger than the target or the remaining weights do not suffice to reach the target, we do not add this sum to the table. If some of the weights are negative, we extend the bounds in both tables, accordingly. Note that the presence of negative weights makes solving the problem more resource-intensive (Daxing and Shaohan 1994). Finally, in order to reduce the size of tables and, hence, the computational requirements, we divide the search into two stages. First, we build the tables with rounded-down weights and find solution candidates at rounded-down precision. Then, we calculate the score of each candidate at the full precision and select the solutions that match the target.

Choosing the best solution

Dynamic programming enables us to find *all* valid genotype combinations that result in a given PRS. We still need a way to select the "best" one among them. The insight that helps us is that variations in the human genome are not uniformly distributed—some genotypes are more likely than others. Hence, we can quantify the likelihood of a genotype for a given variant by its frequency in the population. To simplify the analysis, we consider the SNPs in a PRS model statistically independent from each other and measure the likelihood of a solution by combining the likelihoods of the individual SNP genotypes. Specifically, we calculate the sum of the log-probability for all the solution genotypes based on the allele frequency (AF) in the population that the target individual belongs to and select the solution with the highest sum as the best one. This approach consistently places the true solution in the top ranks of all valid solutions for the majority of individuals. In practice, PRS models might contain correlated variants due to residual linkage disequilibrium. If an attacker identifies such highly correlated variants, they can use them to their advantage by boosting the likelihood of the more probable SNP combination or even avoid adding the unlikely combination to the sums tables, and thereby increasing the prediction accuracy or reducing the computational overhead.

$$P(g_0,\ldots,g_{n-1}) = \sum_{i=0}^{n-1} \log P(g_i|\mathrm{AF}) = \sum_{i=0}^{n-1} \log P(allele_{i_0}, allele_{i_1}|\mathrm{AF}) \quad (3)$$

As the number of SNPs included in a PRS is large, the number of possible solutions grows rapidly, making it impractical to keep track of or to store all the potential solutions. We address this issue with an algorithm that, instead of storing all possible paths, keeps a reference only to the path with the highest likelihood for each sum in the table. The process involves three steps: (1) calculating the likelihood of a solution where all genotypes consist of non-effect alleles, (2) updating the likelihood for each new sum by adjusting the log-probability—subtracting that of the non-effect alleles and adding that of the effect alleles—to the highest likelihood from the previous sum, and (3) keeping a reference to the last added effect weight that maximizes the likelihood for each sum. During backtracking, we trace only the path with the highest likelihood at each step, ultimately identifying the solution with the greatest likelihood. The complete algorithm for recovering genotypes from a PRS is outlined in Supplemental Algorithm S1.

PRS chaining and self-repair

In a realistic scenario, an attacker might have access to multiple PRS values from an individual. Some PRSs share overlapping SNPs, as the same genetic mutation can influence multiple traits. The attacker could exploit this overlap to infer genotypes for these SNPs. The strategy is to first predict genotypes for the PRSs with fewer SNPs, as they can be computed more quickly. Hence, we first solve the PRS with the smallest set of SNPs. For each subsequent PRS that includes more SNPs, we retain the genotypes of overlapping SNPs from the previous solution. This approach gradually reduces the possible solution space as the number of SNPs increases, thus making the computations more efficient.

When inferring genotypes from a single PRS, the attacker lacks a direct way of confirming whether their highest-likelihood solution contains the correct genotypes. However, access to multiple PRSs from the same individual provides a way to cross-check solutions and to increase recovery accuracy. If the attacker cannot find a solution for a PRS using previously inferred genotypes, it likely indicates that some of the earlier predictions were incorrect—for example, the correct solution might have had the second highest likelihood. To address this, we propose a "self-repair" strategy. When a PRS fails to yield a solution with the genotypes predicted from a smaller PRS, the attacker attempts to predict the genotypes without imposing those predictions and generates a sorted list of potential solutions. The attacker then tests each solution in the list, checking whether it enables successful prediction of genotypes from all previously processed PRSs. If a solution validates the genotype prediction for earlier PRSs, the attacker updates their genotypes with this refined solution and proceeds to the next PRS. As genotype solutions are ranked by using population-based likelihoods, incorrect early solutions are therefore unlikely to remain top-ranked as more PRSs are incorporated. This iterative strategy gradually fixes early prediction errors.

As with exploiting residual LD in a single PRS model, our approach can, in theory, be extended to correlated variants across multiple models. In addition to the overlapping SNPs, the attacker can infer the genotypes of correlated SNPs of larger PRSs and thus further reduce the solution search space.

Linking predicted genotypes to databases

Genealogy services enable users to upload their genotype information to the database and search for their potential relatives among other users. The attacker can upload the recovered genotypes of the individual to a genealogy database and discover the genetic matches. The exact algorithms that these services use are not disclosed but prior studies (Ney et al. 2020; Edge and Coop 2020) suggest that they search for long genomic segments with no mismatching homozygous sites between pairs of individuals. Here, we emulate a genealogy search with the individuals from the 1000 Genomes dataset by using popular algorithms for kinship inference from SNPs, namely KING-robust (Manichaikul et al. 2010) and GCTA (Yang et al. 2011).

We first created a database by using the 2,467 unrelated individuals, 25 pairs of first-degree (parent-child, siblings) and 9 pairs of second-degree relatives (grandparent-grandchild, half-siblings, etc.) from the 1000 Genomes data. After predicting ~2,600 genotypes with ~5% error for each individual, we linked them to the emulated database that contained the full set of genotypes for each individual. We used PLINK (Chang et al. 2015; Purcell and Chang 2024) to compute a matrix with pairwise kinship coefficients for both KING-robust (Manichaikul et al. 2010) and GCTA (Yang et al. 2011) algorithms. We then used the relationship inference criteria of 0.354 for self-identification, 0.177 for first-degree relatives, and 0.089 for second-degree relatives (Manichaikul et al. 2010, Table 1). Having the coefficients for all the pairs enabled us to calculate the precision and recall of linking each individual to themselves and their relatives, if the latter were present, and to measure the distribution of the coefficients for each category.

KING-robust (Manichaikul et al. 2010): KING-robust is an algorithm for pairwise relationship inference in the presence of population substructure. The algorithm calculates the kinship coefficient that estimates the probability of two randomly sampled alleles from two individuals being identical by descent (IBD). The high-level idea is that having different homozygous alleles at the same locus reduces the IBD probability. The kinship coefficient is calculated as a relation between the number of loci where two individuals both have heterozygous alleles ($N_{Aa,Aa}$ for reference allele A and alternative allele a), have different homozygous alleles ($N_{AA,aa}$), and the total number of heterozygous alleles (N_{Aa}):

$$\hat{\phi}_{ij} = \frac{N_{Aa,Aa} - 2N_{AA,aa}}{N_{Aa}^{(i)} + N_{Aa}^{(j)}}$$

GCTA (Yang et al. 2011): GCTA is also an algorithm for pairwise estimation of genetic relationships. It measures kinship in terms of how much the genotypes that two individuals have deviate from the expected counts in the population. As it relies on the allele frequency in the population, it is less robust than KING-robust for individuals with different ancestry. Given the number x_k of the reference alleles that an individual has and the frequency p_k of the reference allele in the population for the k^{th} SNP, the kinship between individuals i and j with N SNPs is calculated as follows:

$$\hat{\phi}_{ij} = \frac{1}{2N} \sum_{k=1}^{N} \frac{(x_k^{(i)} - p_k)(x_k^{(j)} - p_k)}{2p_k(1-p_k)}$$

PRS as an identifier to link known individuals to anonymized databases

Given access to an anonymized genomic dataset and the PRS of a known individual, the attacker can compute the PRS for each sample in the dataset. If one matches the target PRS, the attacker could infer the individual's genotypes or associated phenotypes without the need for computationally intensive genotype recovery. This method relies, however, on the assumption that a PRS uniquely identifies the individual. In practice, different genotype combinations can yield the same score, and multiple samples may share an identical set of genotypes for the SNPs in the PRS model.

Here, we demonstrate how an attacker can assess the uniqueness of a given PRS within a population, thereby estimating their confidence that a potential match corresponds to the individual of interest. Intuitively, the uniqueness of a PRS is influenced by factors such as the number of included SNPs, the precision of effect weights (i.e., the number of decimal places), allele frequencies in the population, and the size of the target dataset or population.

PRS as a normal distribution

As genotypes in PRS models are supposed to be independent due to LD clumping, the central limit theorem implies that their linear combination will approximate a normal distribution, given a sufficiently large sample size and a sufficient number of variants in the PRS. This normal distribution is characterized by its mean and standard deviation; therefore, estimating these parameters enables us to determine the probability density associated with any given PRS.

Under the assumption that PRS variants in individuals from mixed non-disease cohorts adhere to the Hardy-Weinberg equilibrium, we can estimate the expected contribution of

variant i to the PRS based on the population allele frequency p_i and the effect weight β_i:

$$\mu_i = 2p_i(1-p_i)\beta_i + p_i^2(\beta_i + \beta_i)$$

The expected mean μ and standard deviation σ of a PRS, given the ploidy P and the number of variants N, are

$$\mu = \sum_{i=1}^{N} \frac{\mu_i}{P \cdot N}$$

$$\sigma = \sum_{i=1}^{N} \frac{\sqrt{(1-p_i)^2(0-\mu_i)^2 + 2p_i(1-p_i)(\beta_i-\mu_i)^2 + p_i^2(2\beta_i-\mu_i)^2}}{P \cdot N}$$

For a PRS to approximate a normal distribution, its effect weights must be densely and evenly distributed. If one weight, for instance, is orders of magnitude larger than the others, it can skew the distribution, potentially creating a trimodal shape. In this study, we observed that most PRSs approximately followed a normal distribution. This is likely because associations with polygenic diseases or traits are typically influenced by multiple variants rather than being dominated by any single one.

Analytical estimation of score uniqueness

Having estimated mean μ and standard deviation σ, we can use the probability density function of the normal distribution to calculate the probability of a score x being in the range $[x-\Delta, x+\Delta]$ as

$$p(x,\mu,\sigma,\Delta) = \int_{x-\Delta}^{x+\Delta} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \tag{4}$$

In order to assess score uniqueness, however, we need the probability of each specific discrete score rather than a continuous range. We reconcile this requirement with the continuity of the normal distribution by defining the probability of a discrete score x_i as the probability of the score being in the range $[x_i-\Delta, x_i+\Delta]$, where $\Delta = \frac{x_{i+1}-x_i}{2}$.

The total number of possible scores and, hence, distance Δ are determined by either the total number of possible combinations of the effect weights or their decimal precision. Given a PRS with N variants and B distinct weights $\{\beta_0, \ldots, \beta_{B-1}\}$, we can calculate the number of possible weight combinations as $L_{comb} = \prod_{i=0}^{B-1} 2 \cdot \text{count}(\beta_i) + 1$, where the function count returns the number of variants that share the same effect weight. We approximate the range of possible scores due to the effect weights' decimal precision for a PRS with N variants as:

$$R = \underbrace{\{\sum_{i=0}^{N-1} 2\beta_i^-\}}_{R_{min}} \cup \underbrace{[\sum_{i=0}^{N-1} 2\beta_i^- - \beta_{max}^-]}_{R_{2min}}, \underbrace{\sum_{i=0}^{N-1} 2\beta_i^+ - \beta_{min}^+]}_{R_{2max}} \cup \underbrace{\{\sum_{i=0}^{N-1} 2\beta_i^+\}}_{R_{max}},$$

where β^+ and β^- are positive and negative effect weights, respectively. We exclude the ranges $[R_{min}, R_{2min}]$ and $[R_{2max}, R_{max}]$ as a PRS cannot take any value from these segments.

The number of possible scores due to the decimal precision becomes

$$L_{prec} = (R_{2max} - R_{2min}) \cdot 10^{\text{max\_num\_decimal}(\beta_0,\ldots,\beta_{N-1})} + 2$$

We can then set $L_{scores} = \min(L_{comb}, L_{prec})$ and calculate $\Delta = \frac{R_{2max}-R_{2min}}{2L_{scores}}$.

Given the score probability defined in Eq. (4) and the calculated Δ, we can estimate the probability of a single sample in a dataset of size M to have score x_i, i.e., being uniquely identifiable, by using the probability mass function of the binomial distribution. In essence, we calculate the probability that M trials of an event with probability p_i result in only one successful outcome.

$$P_{unique}(x_i) = \frac{M!}{1!(M-1)!} \cdot p_i \cdot (1-p_i)^{M-1} = M \cdot p_i \cdot (1-p_i)^{M-1}$$

To calculate the expected total number of unique scores, we sum the probabilities of each score x_i being unique:

$$E_{unique} = \sum_{i=0}^{N-1} M \cdot p_i \cdot (1-p_i)^{M-1}$$

To estimate the median number of samples in the dataset with the same score, we calculate the expected number of samples for each score $x_i \in R$ as $E(x_i) = M \cdot p_i$, store all the results with $E(x_i) \geq 1$ in an array, then sort the array and return the middle element.

Finally, we reduce the computational overhead of the analysis by calculating Eq. (4) with the cumulative distribution function (CDF) of the normal distribution:

$$p_i = \Phi(\frac{x_i + \Delta - \mu}{\sigma}) - \Phi(\frac{x_i - \Delta - \mu}{\sigma})$$

where Φ is the CDF of the normal distribution.

Code Availability

The code and data for running a genotype-recovery demo, plotting the figures from this study, and reproducing the experiments end to end can be found at https://github.com/g2lab/prs-privacy and as Supplemental Code.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

This research was supported by a Warren Alpert Foundation grant to G.G..

Author contributions: G.G. conceived the idea; K.N. developed the methods; G.G. and K.N. designed the experiments; K.N. implemented the software and performed the experiments; G.G. and K.N. analyzed the results, wrote and revised the manuscript.

References

23andMe. 2024. 23andMe launches new genetic reports on common forms of cancer. https://mediacenter.23andme.com/press-releases/23andme-launches-new-genetic-reports-common-forms-cancer. Acc: 2026-06-18.

Aksit MA, Pace RG, Vecchio-Pagán B, Ling H, Rommens JM, Boelle PY, Guillot L, Raraigh KS, Pugh E, Zhang P, et al.. 2020. Genetic modifiers of cystic fibrosis-related diabetes have extensive overlap with type 2 diabetes and related traits. *The Journal of Clinical Endocrinology & Metabolism* **105**.

Allegrini AG, Selzam S, Rimfeld K, von Stumm S, Pingault JB, and Plomin R. 2019. Genomic prediction of cognitive traits in childhood and adolescence. *Molecular Psychiatry* **24**.

ApoE4Info. 2023. Did Nebula Genomics test and am freaking out — looking for a genetic counselor I can speak with. Acc: 2026-02-11.

Austrin P, Kaski P, Koivisto M, and Nederlöf J. 2015. Subset sum in the absence of concentration. In *Symposium on Theoretical Aspects of Computer Science* (eds. EW Mayr and N Ollinger), volume 30 of *LIPIcs*.

Austrin P, Kaski P, Koivisto M, and Nederlöf J. 2016. Dense subset sum may be the hardest. In *Symposium on Theoretical Aspects of Computer Science* (eds. N Ollinger and H Vollmer), volume 47 of *LIPIcs*.

Bellman R. 1957. *Dynamic Programming*. Princeton University Press.

Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, et al.. 2018. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, and Lee JJ. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**.

Coster MJ, Joux A, LaMacchia BA, Odlyzko AM, Schnorr CP, and Stern J. 1992. Improved low-density subset sum algorithms. *Computational complexity* **2**.

Daxing L and Shaohan M. 1994. Two notes on low-density subset sum algorithm. In *The International Symposium on Algorithms and Computation* (eds. DZ Du and XS Zhang).

Desikan RS, Fan CC, Wang Y, Schork AJ, Cabral HJ, Cupples LA, Thompson WK, Besser L, Kukull WA, Holland D, et al.. 2017. Genetic assessment of age-associated Alzheimer disease risk: Development and validation of a polygenic hazard score. *PLoS medicine* **14**.

Edge MD and Coop G. 2020. Attacks on genetic privacy via uploads to genealogical databases. *Elife* **9**.

Erlich Y, Shor T, Pe'er I, and Carmi S. 2018. Identity inference of genomic data using long-range familial searches.

Science **362**.

Fiume M, Cupak M, Keenan S, Rambla J, de la Torre S, Dyke SO, Brookes AJ, Carey K, Lloyd D, Goodhand P, et al.. 2019. Federated discovery and sharing of genomic data using Beacons. *Nature biotechnology* **37**.

Garey MR and Johnson DS. 1979. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman and Company.

Gibson TM, Karyadi DM, Hartley SW, Arnold MA, Berrington de Gonzalez A, Conces MR, Howell RM, Kapoor V, Leisenring WM, Neglia JP, et al.. 2024. Polygenic risk scores, radiation treatment exposures and subsequent cancer risk in childhood cancer survivors. *Nature medicine* **30**.

Gymrek M, McGuire AL, Golan D, Halperin E, and Erlich Y. 2013. Identifying personal genomes by surname inference. *Science* **339**.

Horowitz E and Sahni S. 1974. Computing partitions with applications to the knapsack problem. *Journal of the ACM* **21**.

Hsu L, Jeon J, Brenner H, Gruber SB, Schoen RE, Berndt SI, Chan AT, Chang-Claude J, Du M, Gong J, et al.. 2015. A model to determine colorectal cancer risk using common genetic susceptibility loci. *Gastroenterology* **148**.

Ibanez L, Dube U, Saef B, Budde J, Black K, Medvedeva A, Del-Aguila JL, Davis AA, Perlmutter JS, Harari O, et al.. 2017. Parkinson disease polygenic risk score is associated with Parkinson disease status and age at onset but not with alpha-synuclein cerebrospinal fluid levels. *BMC neurology* **17**.

Jacobs KB, Yeager M, Wacholder S, Craig D, Kraft P, Hunter DJ, Paschal J, Manolio TA, Tucker M, Hoover RN, et al.. 2009. A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies. *Nature genetics* **41**.

Karp RM. 1972. Reducibility among combinatorial problems. In *Complexity of Computer Computations Symposium* (eds. RE Miller, JW Thatcher, and JD Bohlinger).

Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, Natarajan P, Lander ES, Lubitz SA, Ellinor PT, et al.. 2018. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics* **50**.

Khera AV, Emdin CA, Drake I, Natarajan P, Bick AG, Cook NR, Chasman DI, Baber U, Mehran R, Rader DJ, et al.. 2017. Genetic risk, adherence to a healthy lifestyle, and coronary disease. *New England Journal of Medicine* **376**.

Kloeve-Mogensen K, Rohde PD, Twisttmann S, Nygaard M, Koldby KM, Steffensen R, Dahl CM, Rytter D, Overgaard MT, Forman A, et al.. 2021. Polygenic risk score prediction for endometriosis. *Frontiers in Reproductive Health* **3**.

Lagarias JC and Odlyzko AM. 1985. Solving low-density subset sum problems. *Journal of the ACM* **32**.

Lambert SA, Gil L, Jupp S, Ritchie SC, Xu Y, Buniello A, McMahon A, Abraham G, Chapman M, Parkinson H, et al.. 2021. The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. *Nature genetics* **53**.

Lin Z, Owen AB, and Altman RB. 2004. Genomic research and human subject privacy. *Science* **305**.

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, and Chen WM. 2010. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**.

Marston NA, Kamanu FK, Melloni GE, Schnitzler G, Hakim A, Ma RX, Kang H, Chasman DI, Giugliano RP, Ellinor PT, et al.. 2025. Endothelial cell-related genetic variants identify ldl cholesterol-sensitive individuals who derive greater benefit from aggressive lipid lowering. *Nature Medicine* **31**.

Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, Daly MJ, Bustamante CD, and Kenny EE. 2017. Human demographic history impacts genetic risk prediction across diverse populations. *The American Journal of Human Genetics* **100**.

Mavaddat N, Michailidou K, Dennis J, Lush M, Fachal L, Lee A, Tyrer JP, Chen TH, Wang Q, Bolla MK, et al.. 2019. Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *The American Journal of Human Genetics* **104**.

Meyer MN, Papageorge NW, Parens E, Regenberg A, Sugarman J, and Thom K. 2024. Potential corporate uses of polygenic indexes: Starting a conversation about the associated ethics and policy issues. *The American Journal of Human Genetics* **111**.

Musliner KL, Mortensen PB, McGrath JJ, Suppli NP, Hougaard DM, Bybjerg-Grauholm J, Bækvad-Hansen M, Andreassen O, Pedersen CB, Pedersen MG, et al.. 2019. Association of polygenic liabilities for major depression, bipolar disorder, and schizophrenia with risk for depression in the Danish population. *JAMA psychiatry* **76**.

Nebula Genomics. 2024. Nebula reporting, what's inside? https://nebula.org/blog/a-look-at-our-reporting/. Acc: 2024-09-26.

Ney P, Ceze L, and Kohno T. 2020. Genotype extraction and false relative attacks: Security risks to third-party genetic genealogy services beyond identity inference. In *Network and Distributed System Security Symposium*.

Nyholt DR, Yu CE, and Visscher PM. 2009. On Jim Watson's APOE status: genetic information is hard to hide. *European Journal of Human Genetics* **17**.

Pain O, Al-Chalabi A, and Lewis CM. 2024. The GenoPred pipeline: a comprehensive and scalable pipeline for polygenic scoring. *Bioinformatics* **40**.

Pakstis AJ, Speed WC, Fang R, Hyland FC, Furtado MR, Kidd JR, and Kidd KK. 2010. SNPs for a universal individual identification panel. *Human genetics* **127**.

Purcell S and Chang C. 2024. Plink 2.0. Available at: www.cog-genomics.org/plink/2.0/.

Radziszowski S and Kreher D. 1988. Solving subset sum problems with the L^3 algorithm. *The Journal of Combinatorial Mathematics and Combinatorial Computing* **3**.

Ram N and Roberts JL. 2019. Forensic genealogy and the power of defaults. *Nature Biotechnology* **37**.

Reddit. 2021a. Help! Nebula results. Acc: 2026-02-11.

Reddit. 2021b. What am i supposed to make of a polygenic risk score at the 100th percentile? Acc: 2026-02-11.

Saklatvala JR, Lessard S, Teder-Laving M, Thomas LF, Ramessur R, Zierer J, Åsvold BO, Barton A, Baudry D, Bowes J, et al.. 2026. Genetic liability to psoriasis predicts severe disease outcomes. *Genome medicine* **18**.

Schroeppel R and Shamir A. 1981. A $T = O(2^{n/2})$, $S = O(2^{n/4})$ algorithm for certain NP-complete problems. *SIAM Journal on Computing* **10**.

Seibert TM, Fan CC, Wang Y, Zuber V, Karunamuni R, Parsons JK, Eeles RA, Easton DF, Kote-Jarai Z, Al Olama AA, et al.. 2018. Polygenic hazard score to guide screening for aggressive prostate cancer: development and validation in large scale cohorts. *BMJ* **360**.

Tanner A. 2017. *Our bodies, our data: how companies make billions selling our medical records*. Beacon Press.

The 1000 Genomes Project Consortium. 2015a. A global reference for human genetic variation. *Nature* **526**.

The 1000 Genomes Project Consortium. 2015b. IGSR: The International Genome Sample Resource. https://www.internationalgenome.org/category/phase-3/. Acc: 2024-09-18.

The Guardian. 2023. Private UK health data donated for medical research shared with insurance companies. https://www.theguardian.com/technology/2023/nov/12/private-uk-health-data-donated-medical-research-shared-insurance-companies. Acc: 2024-09-18.

Thompson DJ, Wells D, Selzam S, Peneva I, Moore R, Sharp K, Tarran WA, Beard EJ, Riveros-Mckay F, Giner-Delgado C, et al.. 2022. UK Biobank release and systematic evaluation of optimised polygenic risk scores for 53 diseases and quantitative traits. *medRxiv* .

Torkamani A, Wineinger NE, and Topol EJ. 2018. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics* **19**.

Wang R, Li YF, Wang X, Tang H, and Zhou X. 2009. Learning your identity and disease from research papers: information leaks in genome wide association study. In *ACM Conference on Computer and Communications Security*.

Wray NR, Ripke S, Mattheisen M, Trzaskowski M, Byrne EM, Abdellaoui A, Adams MJ, Agerbo E, Air TM, Andlauer TM, et al.. 2018. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature genetics* **50**.

Xu L, Gan T, Chen P, Liu Y, Qu S, Shi S, Liu L, Zhou X, Lv J, and Zhang H. 2024. Clinical application of polygenic risk score in IgA nephropathy. *Phenomics* **4**.

Yanes T, Tiller J, Haining CM, Wallingford C, Otlowski M, Keogh L, McInerney-Leo A, and Lacaze P. 2024. Future implications of polygenic risk scores for life insurance underwriting. *NPJ genomic medicine* **9**.

Yang J, Lee SH, Goddard ME, and Visscher PM. 2011. GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics* **88**.