Assessing transcriptomic reidentification risks using discriminative sequence models

Shuvom Sadhuka,[1,2] Daniel Fridman,[2,3] Bonnie Berger,[1,2] and Hyunghoon Cho[2]

[1]*Computer Science and AI Lab, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA;* [2]*Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA;* [3]*Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts 02115, USA*

Gene expression data provide molecular insights into the functional impact of genetic variation, for example, through expression quantitative trait loci (eQTLs). With an improving understanding of the association between genotypes and gene expression comes a greater concern that gene expression profiles could be matched to genotype profiles of the same individuals in another data set, known as a linking attack. Prior works show such a risk could analyze only a fraction of eQTLs that is independent owing to restrictive model assumptions, leaving the full extent of this risk incompletely understood. To address this challenge, we introduce the discriminative sequence model (DSM), a novel probabilistic framework for predicting a sequence of genotypes based on gene expression data. By modeling the joint distribution over all known eQTLs in a genomic region, DSM improves the power of linking attacks with necessary calibration for linkage disequilibrium and redundant predictive signals. We show greater linking accuracy of DSM compared with existing approaches across a range of attack scenarios and data sets including up to 22,288 individuals, suggesting that DSM helps uncover a substantial additional risk overlooked by previous studies. Our work provides a unified framework for assessing the privacy risks of sharing diverse omics data sets beyond transcriptomics.

[Supplemental material is available for this article.]

The growing availability of large-scale genomic data repositories has led to increasing concerns for the privacy of the individuals from whom the data were collected (Erlich and Narayanan 2014; Naveed et al. 2015; Bonomi et al. 2020). Although many nations and organizations have introduced policies and regulations (e.g., HIPAA and GDPR) to safeguard the collection, use, and sharing of personally-identifying information, existing policies often fall short of providing clear guidance regarding many widely used types of biomedical data, including genetic sequences and functional genomic data, for which the underlying privacy risks are often unclear and only beginning to be understood (Clayton et al. 2019; Gürsoy et al. 2020, 2021, 2022b; Wan et al. 2022; Hill et al. 2023). This lack of guidance, particularly for functional genomic data, presents a key challenge for ensuring the protection of study participants, leaving the possibility for future privacy breaches that may diminish public trust in the scientific community.

Transcriptomic data, such as gene expression measurements broadly shared in databases such as the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) (Barrett et al. 2010) and ArrayExpress (Athar et al. 2019), are a prominent example of biomedical data with incompletely understood privacy implications. Although prior works (Schadt et al. 2012; Harmanci and Gerstein 2016) have shown that the knowledge of *expression quantitative trait loci* (eQTLs)—that is, genetic variants correlated with expression levels of a gene (referred to as eGenes)—could be used to extract genotypic information from gene expression profiles, the full extent of such leakage largely remains unknown. A key concern is that a malicious actor may exploit this leakage to carry out a *linking attack*, in which one links an individual's gene expression profile to their genotype profile in another data set (or vice versa), potentially reidentifying the individual's data and subsequently associating it with a sensitive attribute such as disease status.

Linking attacks require some criterion, which we call a match score function, to score and rank possible candidate matches between the gene expression and genotype data sets. The core challenge in accurately assessing the risk of a linking attack is therefore in devising the best possible match score function that could be leveraged by an attacker. Existing proposals for this function suffer from the key limitation that they account for only a small subset of (nearly) independent eQTLs as required by the simplified probabilistic models used in those approaches (Schadt et al. 2012; Harmanci and Gerstein 2016). As we will show, such restrictions have thus far obscured the extent of genotypic information leakage in expression data.

Here, we introduce the *discriminative sequence model* (DSM), which jointly models all known eQTLs in a genomic region to enable *sequence-level* inference of genotypes given a gene expression profile. DSM provides more accurate probability estimates for each candidate genotype profile originating from the same individual as the expression profile by calibrating for both correlation among genetic variants and redundant eQTL association signals. Using these probabilities as match scores, DSM leads to substantially greater success in simulated linking attacks than existing strategies in a wide range of scenarios, including (1) linking an expression profile against a large-scale candidate genotype set (e.g., including 22,288 individuals) representing a population whose ancestry background differs from the training set, (2) linking when the membership of the target individual in the genotype

Corresponding authors: bab@mit.edu, hhcho@broadinstitute.org
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.277699.123. Freely available online through the *Genome Research* Open Access option.

set is unknown, and (3) linking in the reverse direction, namely, from a genotype profile to a panel of candidate expression profiles. Our work provides an essential tool for gaining a deeper understanding of privacy risks of transcriptomic data and their linkage with protected genetic information.

Results

Overview of the DSM

We developed DSM based on the framework of conditional random fields (Sutton et al. 2012) to model the conditional distribution over the genotype sequence given a query gene expression profile. DSM facilitates linking of samples between gene expression and genotype data sets by scoring the likelihood of a target genetic sequence belonging to the same individual as the query gene expression profile (Fig. 1A,B). DSM extends the widely used Li–Stephens hidden Markov model (HMM) of genetic sequences (Li and Stephens 2003) to include additional probabilistic factors that capture the correlation between each eQTL and its corresponding set of known genes whose expression is correlated with an eQTL (eGenes). In contrast to existing models, DSM obtains calibrated probabilities accounting for all known eQTLs and their genotypic correlations (Fig. 1C,D). We detail our problem setting, threat model, and existing approaches in the Methods.

DSM incorporates a number of novel strategies for joint modeling of genotypes and gene expression (Methods). First, instead of defining a *generative* distribution over gene expression given genotypes (Schadt et al. 2012; Harmanci and Gerstein 2016), we take a *discriminative* approach by parameterizing the distribution over the unknown genotypes given the observed gene expression. This obviates the need for restrictive modeling assumptions over the continuous gene expression space (e.g., normality of expression levels), instead allowing us to work with discrete genotype distributions. Second, DSM learns to calibrate the predictive probabilities during training to adjust for linkage disequilibrium among nearby genetic variants as well as redundant predictive signals across different eQTLs and eGenes. This feature allows the model to leverage the full range of information captured by eQTLs without being limited to a filtered set of independent eQTLs. Third, DSM introduces eQTL probabilistic factors that are generalized to include any number of eGenes for each eQTL rather than separately considering each eQTL–eGene pair. When considering genes with correlated expression levels, combining information across a set of genes can enhance the quality of the predictive signal. Lastly, we developed an efficient haplotype-based approximate inference scheme for DSM to enable the use of the sufficiently large reference panels (which are used to capture genotype correlations) required for accurate prediction.

Overview of our experiments

To evaluate the DSM, we compared the success of linking attacks based on DSM against that of two published Bayesian linking strategies, which we refer to as Gaussian naive Bayes (GNB) (Schadt et al. 2012) and extremity-based linking (EBL) (Harmanci and Gerstein 2016), described in the Methods. We trained each model on pairs of genotype and gene expression profiles from the Genotype-Tissue Expression (GTEx) data set (The GTEx Consortium 2015), which is one of the largest available data sets for eQTL studies that could be leveraged by a potential attacker. The set of eQTLs used by the models are obtained from GTEx, representing an attacker who identifies eQTLs in the available training data. Only the reported loci of eQTLs were used; any distributional parameters were estimated during training. For DSM, we additionally used 1000 haplotypes from the Haplotype Reference Consortium (HRC) data set (The Haplotype Reference Consortium 2016), excluding any overlap with GTEx, for the reference panel underlying the HMM component of DSM (Methods).

We then used the trained models to attempt a linking attack on a separate, nonoverlapping data set, including both genotype and gene expression profiles, shuffled to obscure the links between the two data types. Our primary evaluation setting emulates an attack scenario in which the attacker holds a gene expression profile of interest and tries to identify a genotype profile from the same individual from a pool of candidate genotype profiles in another data set. To this end, we independently assigned the best-matching genotype profile to each gene expression profile using each

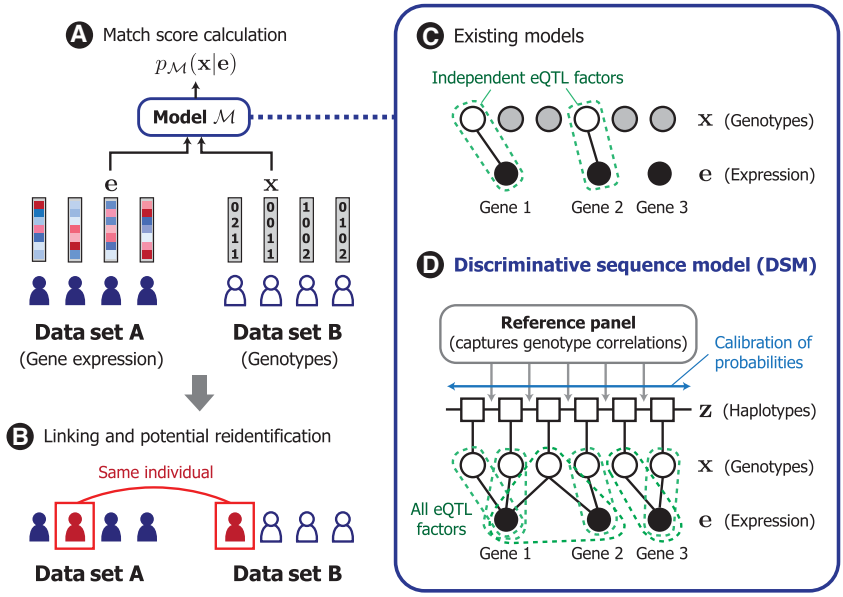

Figure 1. Overview of DSM. (*A*) In our model, we consider the design of a match score function that quantifies how likely a given pair of gene expression and genotype profiles originated from the same individual. (*B*) Such a function can be used by a malicious actor to link individuals across different data sets, which could lead to the reidentification of a data sample corresponding to the individual. (*C*) Existing works investigating this possibility analyzed only a subset of eQTLs (i.e., unshaded nodes, genetic variants associated with gene expression levels) that are statistically independent owing to model limitations. (*D*) We introduce the discriminative sequence model (DSM), which builds upon the standard Li–Stephens hidden Markov model of genetic sequences to jointly leverage the predictive signals across all known eQTLs; it incorporates necessary calibration for redundancy and correlation among eQTLs to provide a more accurate, sequence-level match score. Our modeling approach helps reveal the full extent of genotypic information in gene expression profiles to better inform privacy risk assessment.

Table 1. DSM links more individuals in GTEx cross-validation data sets

Method	Candidate set	Chr 19	Chr 20	Chr 21	Chr 22	All four chromosomes
GNB	GTEx holdout	125 (90.6%)	41 (29.7%)	33 (23.9%)	73 (52.9%)	136 (98.6%)
EBL	GTEx holdout	126 (91.3%)	43 (31.1%)	34 (24.6%)	**76 (55.0%)**	134 (97.1%)
DSM	GTEx holdout	**135/138 (97.8%)**	**87 (63.0%)**	**65 (47.1%)**	60 (43.5%)	**138 (100%)**

The training set and test set are sampled without overlap from the GTEx data set. DSM links every individual in the test set (138) when combining all four chromosomes, as well as a larger fraction of individuals on all but one chromosome individually. Boldface text highlights our method (DSM) and the best linking accuracy in each column.

method and measured the overall accuracy of the attack by the number of individuals for whom we could correctly link their gene expression profiles to their genotypes. As we show below, we considered a range of attack scenarios and different choices of the candidate set to assess the robustness of our method. We evaluated the methods on Chromosomes 19 to 22.

Note that, after observing EBL's poor performance in our experimental setting (owing to relatively few SNPs per chromosome satisfying EBL's extremity thresholds; see Methods), we instead analyzed our extension of EBL, where EBL collapses to GNB for genes with nonextreme expression values. We observed that this hybrid model outperforms the original EBL model, which relies only on predicted genotypes based on extreme gene expression values (see Supplemental Fig. S1). Hyperparameters for EBL and GNB such as the extremity threshold and eQTL–eGene correlation threshold are optimized via a grid search, and the optimal choice of parameters was used to consider the best-case performance for each method.

DSM links more individuals than previous approaches

We first considered the setting in which the training and test data sets are both sampled from the same data set cohort. Such a scenario may occur when an attacker obtains access to training data consisting of individuals from the same population as the target individuals. For this analysis, we used a subset of gene expression and genotype profile pairs of 450 individuals in the GTEx muscle tissue data set (The GTEx Consortium 2015) for training and tested the linking accuracy on a nonoverlapping, held-out set of 138 individuals also from GTEx.

DSM was able to link all 138 test individuals using all four chromosomes and 135 using Chromosome 19 alone (Table 1). Although GNB and EBL also correctly link most individuals (136 and 134 out of 138, respectively) using all four chromosomes, when using a single chromosome, DSM's linking accuracy is greater for all but one chromosome tested (22), suggesting that DSM is able to make better use of limited predictive signals. The performance gap is especially pronounced for Chromosomes 20 and 21: for Chromosome 20, GNB, EBL, and DSM correctly linked 41, 43, and 87 individuals, respectively.

DSM enhances cross–data set linking accuracy

We next assessed our method in the setting in which there is a mismatch between the populations of the training data set and the target data set. Here, we used all 588 pairs of gene expression and genotype profiles available in GTEx for training. For the test data set, we used gene expression and genotype profiles of 292 individuals in the Finland–US Investigation of NIDDM Genetics (FUSION) data set (Ghosh et al. 2000). The gene expression data in FUSION are also obtained from muscle tissue. Notably, there is some ancestry mismatch between GTEx and FUSION, because individuals in the GTEx data set were recruited in the United States and thus were expected to have limited representation of Scandinavian ancestry, in contrast to FUSION, which includes only Finnish individuals. We used the same eQTL set and the HRC reference panel for DSM as the previous analysis, except the latter now excludes any overlap with both GTEx and FUSION.

Using DSM, we were able to link 289 of 292 individuals from Chromosome 19 alone (Table 2). Compared with that of GNB and EBL, our improvement is most clear on chromosomes with fewer eQTLs, such as Chromosome 21 (fewest eQTLs of four chromosomes, including around 11,000), where the DSM correctly links 110 individuals and GNB and EBL link 34 and 36 individuals, respectively. All three models are able to link >99% of individuals when combining all four chromosomes, but DSM provides greater linking accuracy on all four chromosomes individually.

To evaluate the ability of each method to distinguish the true match from a larger set of candidates, we next added 292 additional genotype profiles (584 phased haplotypes) from the GOT2D consortium data (Flannick et al. 2019), of which FUSION is a subset, to the set of candidate genotype profiles for linking. In this scenario, we expect the set of individuals still correctly linked to be a subset of those linked in the previous experiment based only on

Table 2. DSM enhances linking accuracy in a test population different from the training data set

Method	Candidate set	Chr 19	Chr 20	Chr 21	Chr 22	All four chromosomes
GNB	FUSION	247 (84.6%)	74 (25.3%)	34 (11.6%)	162 (55.5%)	290 (99.3%)
EBL	FUSION	255 (87.3%)	80 (27.4%)	36 (12.3%)	166 (56.8%)	290 (99.3%)
DSM	FUSION	**289/292 (99.0%)**	**188 (64.4%)**	**110 (37.7%)**	**241 (82.5%)**	**292 (100%)**
GNB	FUSION + GOT2D	224 (76.7%)	56 (19.2%)	17 (5.8%)	99 (33.9%)	286 (97.9%)
EBL	FUSION + GOT2D	234 (80.1%)	63 (21.6%)	16 (5.5%)	111 (38.0%)	289 (99.0%)
DSM	FUSION + GOT2D	**280/292 (95.9%)**	**88 (30.1%)**	**52 (17.8%)**	**143 (49.0%)**	**292 (100%)**

When linking individuals in the FUSION data set with models trained on the GTEx data set, we are able to link more individuals than previous methods, with the exception of one chromosome. When we add an additional 292 target genomes from the GOT2D consortium to the candidate set (FUSION + GOT2D), our accuracy improvement becomes more pronounced. Boldface text highlights our method (DSM) and the best linking accuracy in each column, separately for each candidate set considered.

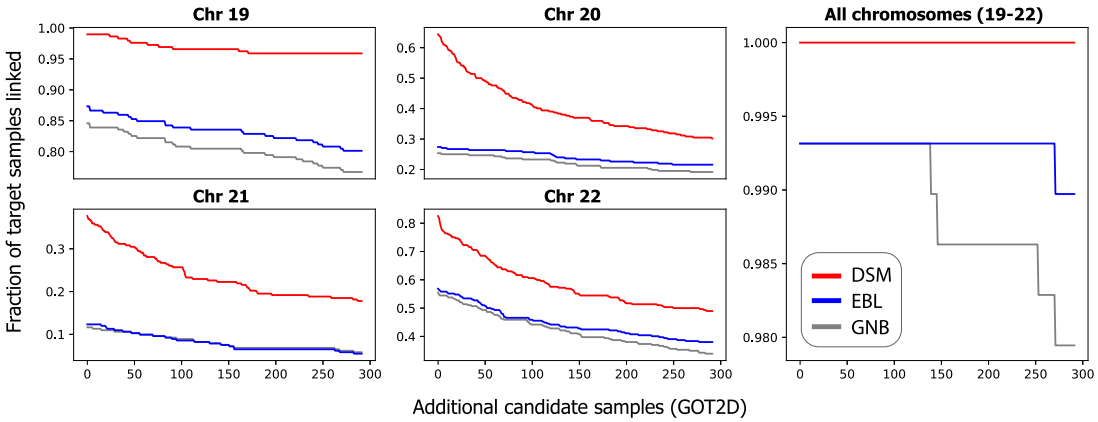

Figure 2. Linking based on DSM remains accurate with larger candidate genotype sets. The plots depict how the linking accuracy of each method for FUSION individuals changes as more nonmatching candidate target genotypes from the GOT2D consortium are added. Results are shown for each of Chromosomes 19–22 and all four chromosomes combined to show performance on different sets of eQTLs. We add the same set of additional candidate samples for each chromosome. Each step corresponds to one individual unlinked. DSM remains accurate for most target individuals, even as more candidate genotype profiles are added.

FUSION, because increasing the number of candidates only makes the match score of the true link less likely to be the best score.

Our results show that DSM is still able to link all 292 individuals combining the four chromosomes (Table 2; Fig. 2). The accuracy of GNB and EBL was substantially reduced by the additional candidates, for example, resulting in 17 and 16 correctly linked individuals, respectively, for Chromosome 21 compared with 52 by DSM. This drop in linking accuracy is observed even for a modest number of added samples (Fig. 2). With just 100 additional samples, GNB and EBL newly miss at least 10 individuals on three of the four chromosomes.

DSM extracts stronger identifying signals from gene expression

To gain a deeper insight into DSM's robust performance on larger sets of candidate genotype profiles, we aimed to assess the *gap* in the match score between the true match and the remaining mismatching pairs. Intuitively, if the gap is large, then we should expect the model to remain accurate for longer as additional samples are added to the candidate set, as it would be less likely that random samples of other individuals' genotype profiles will result in match scores as extreme as the true match.

To assess this gap, we modeled the null distribution of match scores for each of the three models (see Supplemental Note S1), which allowed us to compute the *P*-value representing the strength of the identifying signal of each matching $\mathbf{x}^{(i)}$ and $\mathbf{e}^{(i)}$. Although all three models were able to separate the matching pair from the mismatching pair for >99% of individuals in FUSION (Table 2) with low *P*-values, DSM was able to generate a lower *P*-value for 229 (78.4%) and 239 (81.8%) individuals compared with GNB and EBL, respectively, suggesting greater separation of the true match (Fig. 3A–C).

This result explains our stronger performance when adding extraneous GOT2D genomes; one can imagine adding new GOT2D genomes as sampling from the null match score distribution, so if the *P*-value of the correct match is lower, more samples must be generated to find a mismatching pair with a better score than the true match. We observed lower *P*-values for DSM also based on a larger candidate set including GOT2D individuals (see Supplemental Fig. S2). These results indicate that DSM more sharply separates the matching pair from mismatching pairs.

Finding the needle in a haystack: DSM enables linking against massive candidate genotype sets

We set out to test whether a linking attack based on the DSM is feasible even on a large-scale candidate genotype set including tens of thousands of individuals. This represents a plausible attack scenario in which the attacker takes a target gene expression profile and searches against one of a growing number of large biobank-scale genomic data sets in which the individual may be known to be included. In such a scenario, a large number of nonmatching genotype profiles must be distinguished from the true match in order to obtain a correct link.

To this end, we selected a total of 21,996 genotype profiles (43,992 phased haplotypes) from HRC that did not overlap with FUSION, GTEx, or the DSM reference panel and additionally included this cohort as part of the candidate set together with the FUSION individuals. We observed that DSM displays remarkable robustness to such large data sets, linking 260 of 292 FUSION individuals (89.0%) compared with just 129 (44.1%) and 147 (50.3%) by GNB and EBL, respectively, for Chromosome 20 (Table 3; Fig. 4B). When combining all four chromosomes, DSM linked 273 (93.5%), notably more than 267 (91.4%) and 260 (89.0%) for EBL and GNB, respectively (Table 3; Fig. 4A). Similar gaps were observed for all chromosomes individually (Supplemental Fig. S3).

Furthermore, testing the linking attack on a much larger candidate set allowed us to validate our model-based *P*-values, calculated on a subsampled candidate set (including 500 individuals) using our model of null distribution (Supplemental Note S1), by comparing them to the empirical *P*-values calculated on the full candidate set (Fig. 4C). Increasing noise is expected at the tail of the distribution given the large sample size requirements for empirically estimating small *P*-values. Nevertheless, the high concordance between the model-based and empirical *P*-values (Pearson R^2 of 0.89, log-transformed) provides further evidence that the enhanced identifying signals obtained by DSM (as measured by the *P*-values) truly reflect more robust linking performance in general.

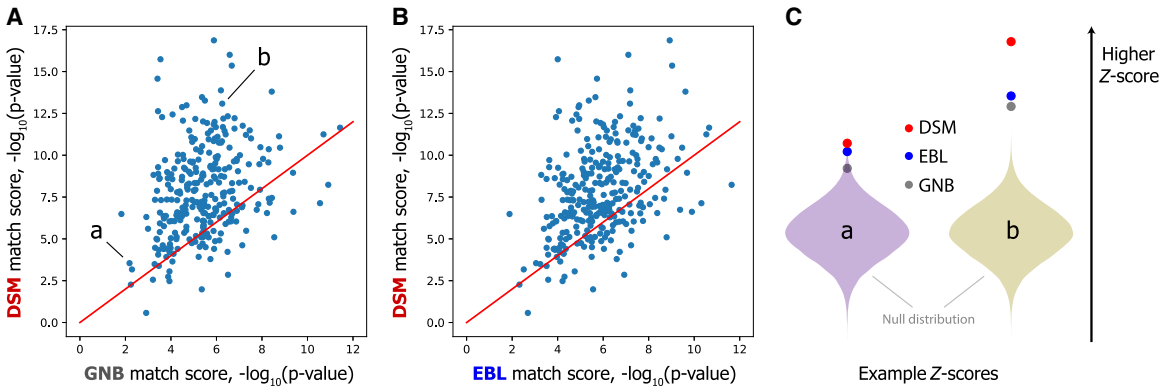

Figure 3. DSM distinguishes matching pairs of profiles from mismatching pairs with higher confidence. (*A*,*B*) DSM separates most individuals in the test data set from the background distribution of match scores more distinctly than both GNB (*A*) and EBL (*B*), as assessed by the *P*-value of the true match score compared with null distribution of mismatching pairs' scores (FUSION individuals only). In *C*, we visualize the strength of separation for the two individuals marked (a and b) in relation to the null distribution normalized as a Gaussian distribution. The DSM indeed separates these individuals more than GNB and EBL. Note that individual a is the one individual that all three methods fail to link, and this individual is substantially less separated from the null distribution than the other individuals.

Additional linking attack scenarios

We thus far have evaluated the accuracy of linking attacks when the genotype profile of the same individual as the query gene expression profile is included in the target data set. In an attack scenario in which the membership of the matching individual in the target data set is unknown to the attacker, the attacker must make an additional decision about whether to draw a link between an expression profile and the top-matching genotype profile. For instance, this can be achieved by imposing a threshold on the match scores such that only the high-confidence links above the threshold are called. To assess the performance of DSM in this setting, we performed a holdout experiment based on the expanded FUSION data set (with 22,288 candidate genotype profiles), in which the matching genotypes of half of the query expression profiles (146 out of 292) were excluded from the target set before the linking procedure. For each method, we evaluated the linking results across a range of match score thresholds with respect to the standard performance metrics for binary classification, such as false-positive rate (FPR), precision, and recall, which we adapted for the linking problem (see Supplemental Note S2).

Our results show that linking based on DSM provides a significantly better tradeoff between identifying true matches and minimizing false positives compared with EBL and GNB (Fig. 5). For example, we observed that the FPR (the proportion of query expression profiles without any matching genotypes that were incorrectly linked) of our linking approach is small (<1%) when identifying half of the true matches with top-match scores, indicated by a true-positive rate (TPR) of 50% (Fig. 5A). Naturally, identifying more true matches leads to a higher FPR (e.g., a FPR of 8% for a TPR of 80%). In contrast, EBL and GNB led to substantially higher FPRs for the same TPR; for example, for a TPR of 80%, they obtained a FPR of 35% and 34%, respectively. In agreement with these observations, DSM obtained a greater AUROC metric (0.887) than both EBL (0.809) and GNB (0.799). We observed a similar improvement of DSM in terms of the tradeoff between the precision and recall metrics across different match score thresholds (Fig. 5B).

We next investigated another relevant attack scenario in which the attacker holds a genotype profile and wishes to match to a gene expression profile in the target data set. We refer to this as *reverse* linking to distinguish it from our main attack scenario. In this setting, the attacker could directly leverage an existing method for predicting gene expression from genotypes to compare with the candidate expression profiles. To this end, we compared our method—using the same scores from DSM as before but performing linking in reverse—against linking based on MetaXcan (Barbeira et al. 2018), a state-of-the-art method for gene expression prediction. We considered two different match scores for MetaXcan: Pearson and Spearman's correlation coefficients between the predicted and observed expression profiles. Training the models on the GTEx muscle-skeletal data set and evaluating reverse linking on the FUSION data set, we observed near-perfect linking accuracy for DSM when combining all four chromosomes (291/292), substantially more accurate than MetaXcan-based linking (Pearson: 171/292, Spearman's: 97/292) (see Supplemental Table S1). For individual chromosomes, reverse linking was generally less accurate than our original results in the forward direction based on DSM (Table

Table 3. DSM robustly links individuals in massive candidate genotype sets

Method	Candidate set	Chr 19	Chr 20	Chr 21	Chr 22	All four chromosomes
GNB	FUSION + HRC	129 (44.1%)	4 (1.4%)	1 (0.3%)	27 (9.2%)	260 (89.0%)
EBL	FUSION + HRC	147 (50.3%)	9 (3.1%)	1 (0.3%)	31 (10.6%)	267 (91.4%)
DSM	FUSION + HRC	**260/292 (89.0%)**	**19 (6.5%)**	**11 (3.8%)**	**38 (13.0%)**	**273 (93.5%)**

When the expanded candidate genotype set including the HRC cohort contains approximately two orders of magnitude more individuals (22,288) than the original FUSION data set (292), DSM consistently links more individuals than EBL and GNB across all four chromosomes. Boldface text highlights our method (DSM) and the best linking accuracy in each column.

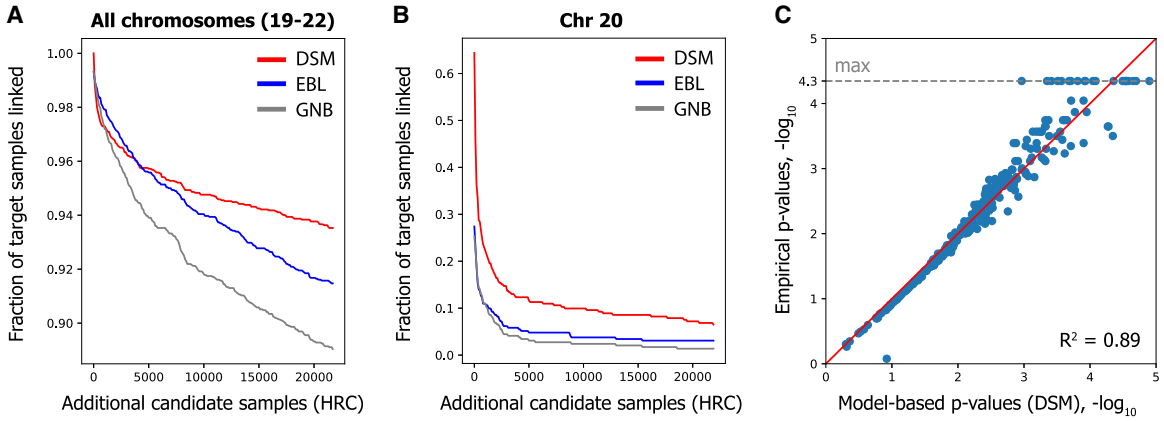

Figure 4. DSM enables linking on massive candidate genotype sets and provides a measure of identifying signals from gene expression data. We included an additional 21,996 HRC individuals in the candidate genotype set to evaluate linking accuracy on massive data sets. (*A*) Combining all four chromosomes, DSM links 94% of individuals. Each curve represents an average across 10 random permutations of the HRC individuals. (*B*) DSM's substantial accuracy improvement is observed for each chromosome, as depicted here for Chromosome 20. Plots for other chromosomes are provided in Supplemental Figure S3. (*C*) A strong correlation is observed between the empirical *P*-values of true matches and our model-based estimates on a random subset of 500 individuals, suggesting that our *P*-values provide a calibrated measure for assessing our method's accuracy on larger candidate sets. The minimum empirical *P*-value (1/22,287), or equivalently maximum negative log *P*-value, is highlighted with the dashed gray line.

2), but DSM still obtained greater linking accuracy than MetaXcan. We hypothesize that the reduced accuracy of reverse linking is owing to the greater impact of noise (both biological and experimental) in gene expression profiles on reverse linking, because a large number of expression profiles are jointly considered to determine the match for each query genotype profile.

Discussion

We have shown that existing models for transcriptomic linking attacks overlook a substantial number of predictive signals captured by eQTLs. Our *discriminative sequence model* (DSM) poses the linking attack as a sequence-level probabilistic modeling problem, one in which the attacker jointly predicts the entire genotype sequence instead of independently predicting each eQTL. Our results show that DSM reveals greater linking attack risk than previous methods over a range of evaluation settings, including cross–data set prediction and linking attacks involving a massive candidate set. We also illustrated how the *P*-values calculated based on the match score distributions can be a proxy for the strength of identifying signals captured by the model, representing a promising new approach for quantifying privacy risks.

Our framework unifies QTL privacy analyses, because it allows joint prediction of genotypes from any source. It would be interesting to explore incorporating other sources of phenotypic data that could leak genotype information into the DSM, including protein abundance (pQTLs) (Hill et al. 2023), methylation (mQTLs) (Gaunt et al. 2016; Backes et al. 2017), and allele-specific expression (aseQTLs) (Gürsoy et al. 2021). Another promising direction would be extending the DSM formulation to the coexpression of genes conditional upon eQTLs, which has been used by other models (Gamazon et al. 2015; Gusev et al. 2016).

There are a few limitations of our work. Training the DSM requires substantial memory (75 GB) and time (~6 h) per window. These requirements depend on the size of the eQTL window and reference panel size (see Supplemental Fig. S4). In contrast, both EBL and GNB can be trained with minimal time and memory requirements. However, once trained, DSM performs matching with far less memory (12 GB) and time (~5 min) than required for training, resulting in less than a day of runtime for the large-scale matching we showed with 22,288 individuals. Also, for assessing the risk of linking attacks, it may be desirable to consider an adversary with large computational resources.

Although our work reveals that gene expression data contain more identifying information than previously known, we acknowledge that further investigation is needed to ascertain the real-world implications of our findings in the context of existing transcriptomic databases. For instance, when the target genotype set does not include the individual associated with the gene expression profile, the risk of a successful linkage is reduced as only the matches with especially high match scores can be called with sufficiently low false detection error, as we illustrated in our experiments (see section "Additional linking attack scenarios"). Furthermore, there are other key factors that modulate the success of potential linking attacks, including ancestry, assay platform for gene expression measurements, and the tissue of origin of the samples. When the target set is different than the training set in these aspects, eQTL-based linking may have reduced accuracy but would still remain feasible (Harmanci and Gerstein 2016). Thus, an appropriate level of protection for gene expression data would ultimately depend on the specific setting of the data set as well as the availability of other public data resources that could be jointly leveraged for an attack.

Beyond standard access control mechanisms, other tools for consideration to enhance data protection include data sanitization (Harmanci and Gerstein 2018; He et al. 2020; Yilmaz et al. 2020; Ye et al. 2022; Zhang and Bonomi 2022; Gürsoy et al. 2022b), differential privacy (Tramèr et al. 2015; Almadhoun et al. 2020; Chen et al. 2021), and secure analysis and storage platforms (Lambert et al. 2018; Dokmai et al. 2021; Froelicher et al. 2021; Cho et al. 2022; Gürsoy et al. 2022a). Investigating the use of these techniques to provide meaningful mitigation of the risks illustrated in our work, while maintaining the scientific utility of sharing genomic and transcriptomic data, is an important next step. All these efforts lay the foundation for secure sharing of omics data.

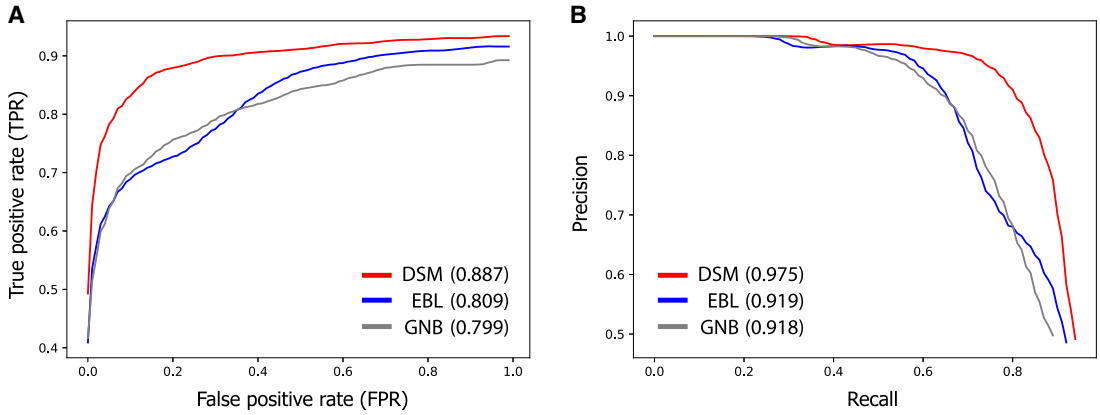

Figure 5. DSM leads to fewer false positives when linking individuals with unknown membership in the target genotype set. When it is unknown whether the individual associated with a query expression profile is included in the target data set, a match score threshold can be used to draw links only for scores greater than the threshold. The choice of this threshold determines the tradeoff between detecting true matches and avoiding spurious links. Receiver-operating-characteristic (ROC) curves (*A*) and precision-recall curves (*B*) illustrating this tradeoff are shown for DSM, EBL, and GNB, evaluated on the expanded HRC data set with 22,288 individuals. The average AUROC and AUPRC metrics are reported in parentheses for each method. Full descriptions of the relevant evaluation metrics (e.g., true-positive rate and false-positive rate) are provided in Supplemental Note S2. The curves are averaged over 100 trials of the holdout experiment in which the genotype profiles of a random half of FUSION individuals were excluded from the target set before each trial of the linking procedure to assess false detection error. DSM outperforms previous methods in finding more true matches while minimizing the number of incorrect links for individuals without a match.

Methods

Problem description and threat model

We consider the setting in which a malicious actor (attacker) wishes to link an individual's gene expression profile to his or her genotype profile in a different data set (Schadt et al. 2012; Harmanci and Gerstein 2016). Both gene expression and genotype data could be either acquired from a public repository or accessed through an authorized channel. In common data sharing scenarios, both types of data are provided without explicit personal identifiers; however, they may include clinical and demographic attributes deemed necessary for the respective studies. These additional attributes, when combined between the two data sources, may lead to the reidentification of individuals even when their identity is not known to the attacker (Sweeney 2000). Studies have also shown that the genetic sequence by itself can be a sufficient identifier when it is searched against public genealogy databases (Gymrek et al. 2013; Erlich et al. 2018). These findings illustrate a plausible threat of reidentification when a gene expression profile is linked to the corresponding genetic sequence.

The goal of the attacker is to use a match score function that, given a query gene expression profile, scores all target genotype profiles as possible links based on some criterion. We consider this function to be the attacker's most important tool, which must be learned properly in order to successfully carry out a linking attack. We thus assume that a motivated attacker will obtain access to a few additional data sets for learning the match score function, including (1) a publicly available list of known eQTL associations (e.g., in the GTEx Portal) (The GTEx Consortium 2015), (2) a training data set of genotype and gene expression profile pairs, and (3) a set of genetic sequences (without associated gene expression data) for modeling genotype distributions. We note that the latter two data sources are commonly available in existing data repositories (e.g., the NCBI database of Genotypes and Phenotypes [dbGaP; https://www.ncbi.nlm.nih.gov/gap/]) (Tryka et al. 2014) and are generally growing in size and availability. Although data use agreements for public repositories typically disallow attempts for reidentification, it is worth noting the attacker could use these additional sources only to train a match score function and not to reidentify individuals in those data sets. Moreover, without clear mechanisms for data provenance and for detection of a breach of agreements, it is plausible that an attacker could use these data sets without repercussions.

For the purpose of comparing the effectiveness of different match scores, we primarily consider the setting in which the attacker knows that a given query gene expression profile matches one of the genotype profiles in the target data set. In practice, the attacker may not know the membership of the individual in the target data set and thus might need to make an additional decision about whether the best possible match found truly represents the same individual (e.g., by imposing a minimum threshold on the match score). In the section "Additional linking attack scenarios," we also address the performance of our method in this setting, as well as for the closely related problem of matching a genotype profile across a set of candidate gene expression profiles (i.e., linking in the opposite direction from that of our main setting).

Notation and definitions

We first provide a formal definition of the transcriptomic linking problem. The attacker has access to two data sets $\mathcal{D}_X$ and $\mathcal{D}_E$, which correspond to a data set of genotypes and gene expression profiles, respectively. We let each $\mathbf{x}^{(j)} \in \mathcal{D}_X$ be a phased genotype profile over biallelic variants (a pair of haplotypes); that is, $\mathbf{x}^{(j)} \in \{(0, 0), (0, 1), (1, 0), (1, 1)\}^V$, where V is the number of variants considered. Any unphased genotype data set can be preprocessed using standard phasing algorithms to obtain a phased data set (e.g., Loh et al. 2016; Delaneau et al. 2019; Browning et al. 2021). Each $\mathbf{e}^{(j)} \in \mathcal{D}_E$ is a vector of gene expression level measurements obtained via microarray or RNA sequencing experiments; that is, $\mathbf{e}^{(j)} \in \mathbb{R}^G$, where G is the number of genes (e.g., around 20,000).

For each individual gene expression profile $\mathbf{e}^{(j)} \in \mathcal{D}_E$, the attacker scores all candidate genotype profiles $\mathbf{x}^{(i)} \in \mathcal{D}_X$ according to some match score function $M(\mathbf{x}^{(i)}, \mathbf{e}^{(j)})$. This function must return a high match score if $\mathbf{x}^{(i)}$ and $\mathbf{e}^{(j)}$ are collected from the same individual; a low score, otherwise. It is natural to use a probabilistic interpretation for M such that $M(\mathbf{x}^{(i)}, \mathbf{e}^{(j)}) = p(\mathbf{x}^{(i)}|\mathbf{e}^{(j)})$

(or $p(\mathbf{e}^{(j)}|\mathbf{x}^{(i)})$), because this represents the probability of observing the target genotype given the query gene expression vector (or vice versa). As previously described, we assume that the attacker trains the match score function M based on an independent training data set $\mathcal{D}' = \{(\mathbf{x}'^{(i)}, \mathbf{e}'^{(i)})\}_{i=1}^{N}$, including matching pairs of genotype and gene expression profiles. With this formulation, obtaining a suitable M thus becomes a statistical learning problem, in which the attacker optimizes the parameters of M to maximize the likelihood $p(\mathbf{x}'^{(i)}|\mathbf{e}'^{(i)})$ over the training data set and then relies on the generalization of M to perform linking between the target data sets $\mathcal{D}_X$ and $\mathcal{D}_E$.

Review of previous linking approaches

Previous works have shown that using a learned model to define the match score function M can provide the ability to predict genotypes based on gene expression (Schadt et al. 2012; Harmanci and Gerstein 2016). These works leverage a set of genetic variants, called eQTLs, whose genotype values are correlated with the expression levels of a gene and thus can be predicted given a gene expression profile.

GNB approach (Schadt et al. 2012)

From here onward, $\mathbf{x}$ (of length V) refers only to the eQTL subset of the genotype profile. Also, let $Q(s) \subset [G]$ denote the set of genes associated with eQTL s, and $\mathbf{e}_{Q(s)}$ denote the corresponding subset of $\mathbf{e}$. Schadt et al. 2012 begin by observing

$$p(\mathbf{x}^{(i)}|\,\mathbf{e}^{(j)}) \propto p(\mathbf{e}^{(j)}|\,\mathbf{x}^{(i)})p(\mathbf{x}^{(i)}), \tag{1}$$

by Bayes' theorem. To simplify the prediction, they further make an *independence* assumption across eQTLs to express $p(\mathbf{x}^{(i)}|\mathbf{e}^{(j)})$ as a product of probabilities for individual eQTLs:

$$p(\mathbf{x}^{(i)}|\mathbf{e}^{(j)}) = \prod_{s=1}^{V} p(\mathbf{x}_s^{(i)}|\mathbf{e}_{Q(s)}^{(j)}) \propto \prod_{s=1}^{V} p(\mathbf{e}_{Q(s)}^{(j)}|\mathbf{x}_s^{(i)}) \cdot p(\mathbf{x}_s^{(i)}). \tag{2}$$

The final expression is used as the match score with additional heuristics for normalization. Both probability terms in the score are estimated on the training data: The conditional distribution over gene expression $p(\mathbf{e}_{Q(s)}|\mathbf{x}_s)$ (with a singleton $Q(s)$ in the setting of Schadt et al. 2012) is parameterized as a Gaussian distribution $\mathcal{N}(\mu_s, \sigma_s)$ with mean μ_s and standard deviation σ_s, whereas the prior distribution $p(\mathbf{x}_s^{(i)})$ is determined by the genotype frequencies. We refer to this approach as Gaussian naive Bayes (GNB), given the Gaussian conditional distribution and the factorization based on independence assumption, which is analogous to naïve Bayes classifiers.

EBL approach (Harmanci and Gerstein 2016)

Given the noisy nature of gene expression measurements, the strongest predictive signals often lie in the extreme values of gene expression levels, for which a Gaussian model may not lead to reliable probability estimates. Harmanci and Gerstein (2016) leverage this insight to develop an alternative linking strategy that directly uses extreme gene expression values for prediction. Specifically, they set $p(\mathbf{x}_s^{(i)}|\mathbf{e}_{Q(s)}^{(j)}) = 1$ (considering only singleton $Q(s)$), for $\mathbf{x}_s^{(i)} = (0, 0)$ or $(1, 1)$ depending on the direction of association, when (1) $\mathbf{e}_{Q(s)}^{(j)}$ is sufficiently "extreme" according to its empirical distribution and (2) $\mathbf{x}_j$ and $\mathbf{e}_{Q(s)}$ are strongly correlated, for suitable thresholds for extremity and correlation. The match score is then computed by comparing the predicted genotypes and the given candidate genotype profile.

Limitations and other related works

Both GNB and EBL approaches treat each SNP independently. To ensure that this assumption is justifiable, the set of eQTLs used by previous models is pruned to a smaller set of mutually independent eQTLs. However, this approach omits all other eQTLs, which collectively provide far richer information about the genotypes than a pruned set would provide, as we show in our work. Simply including all eQTLs in the prediction, disregarding the correlations and redundant predictive signals, lead to poor predictive performance owing to miscalibrated probabilities (see Supplemental Fig. S5), which necessitates our new modeling approaches.

Other works have explored reidentification attacks based on sparse and noisy genotypes (e.g., those obtained from a coffee cup), where a HMM is used to infer matches while considering the haplotype structure (Emani et al. 2021). These works leverage the key insight that higher-order correlations (LD) exist between SNPs and that recombination modeling enables more accurate privacy risk estimation in these scenarios (Samani et al. 2015; Deznabi et al. 2018). Although our work is motivated by a similar insight, unlike the existing works, we introduce an end-to-end learning framework for sequence-level prediction of genotypes based on weak statistical signals from nongenotypic data and show the effectiveness of this approach on gene expression data. Furthermore, acquiring calibrated probabilities for target genotypes is a challenging, yet important task for risk quantification and is uniquely addressed by our work.

Our novel approaches for predicting genotypes from gene expression

Here, we summarize the modeling techniques we newly leverage to improve upon the existing models of genotype prediction based on gene expression.

Discriminative probabilistic modeling

Recall that our goal is to learn a match score function $M(\mathbf{x}, \mathbf{e})$ in terms of $p(\mathbf{x}|\mathbf{e})$ to obtain a ranking over candidate genotype profiles $\mathbf{x}$ given a particular expression profile vector $\mathbf{e}$. As we described, existing methods use the relation $p(\mathbf{x}|\mathbf{e}) \propto p(\mathbf{e}|\mathbf{x})p(\mathbf{x})$ to instead model the conditional distribution over expression given genotypes. This is considered a **generative** approach to probabilistic modeling, in which the prediction is made by hypothesizing different values of the unobserved variable ($\mathbf{x}$) and choosing one that most likely has generated the observed data ($\mathbf{e}$). However, because $\mathbf{e}$ is high-dimensional and has a continuous domain, any parameterization of $p(\mathbf{e}|\mathbf{x})$ must make strong assumptions about its distribution (e.g., normality), which most likely introduces inaccuracies. Also, when multiple genes are associated with an eQTL, modeling the joint distribution $p(\mathbf{e}_{Q(s)}|\mathbf{x}_s)$ is expected to be even more challenging; this, in part, explains why previous works only considered single-gene settings. Instead, we adopt a **discriminative** approach, whereby the target distribution $p(\mathbf{x}|\mathbf{e})$ is directly parameterized and learned. This obviates the need to model the distribution over $\mathbf{e}$, significantly simplifying the problem. In addition, $\mathbf{x}$ is a vector of discrete values, which requires fewer parametric assumptions.

Capturing genotype correlations

Previous models assumed that $p(\mathbf{x}\,|\,\mathbf{e})$ could be factorized into independent probabilities for each eQTL, that is, $\prod_s p(\mathbf{x}_s\,|\,\mathbf{e}_{Q(s)})$, which allows each term to be estimated and calculated independently. However, this requires pruning of correlated eQTLs, resulting in

a smaller set of eQTLs to use for prediction and obscuring the true extent of genotypic information leakage. To address this challenge, we incorporate the Li and Stephens' model of haplotype sequences (Li and Stephens 2003), based on a HMM, as the backbone of our probabilistic model for $p(\mathbf{x}|\mathbf{e})$. Note that the HMM defines a probability distribution over a haplotype sequence, viewing it as a mosaic copy of a collection of haplotype sequences in a reference panel. At each genomic position, there is a probability of crossing over to a different reference haplotype to copy from, which represents the recombination process. This model component thus captures the correlation (i.e., linkage disequilibrium) structure among nearby eQTLs, allowing us to make predictions that correctly account for this structure while considering all eQTLs.

Going beyond single-gene predictions

Another shortcoming of previous models is that they could only consider single eQTL–eGene pairs at a time, that is, $p(\mathbf{e}_{Q(s)}|\mathbf{x}_s)$ with a singleton $Q(s)$. In reality, multiple genes can be associated with a particular eQTL, and the full set of associated genes may provide richer and more accurate information about the genetic variant than only considering the most significant gene. For example, one could reduce the experimental noise in gene expression data by averaging information across correlated genes. Our model allows any number of genes to be used in calculating the predictive signal for a particular eQTL, captured by the following probabilistic factor in our model, instantiated for each eQTL s:

$$\phi_s(\mathbf{x}_s,\ \mathbf{e}_{Q(s)};\ \boldsymbol{\alpha}_s,\ \boldsymbol{\beta}_s) = \sigma(\mathbf{x}_s;\ \boldsymbol{\alpha}_s,\ \boldsymbol{\beta}_s^T\mathbf{e}_{Q(s)}). \tag{3}$$

This component allows us to include an arbitrary number of genes in the partial prediction of $\mathbf{x}_s$ (further refined through the global model) without overly increasing model complexity.

Discriminative sequence model (DSM)

We newly introduce DSM, a probabilistic graphical model that enables joint prediction of $\mathbf{x}$ from $\mathbf{e}$ by combining the insights described above. More precisely, we model the conditional distribution $p(\mathbf{x}|\mathbf{e})$ as a conditional random field (CRF) (Sutton et al. 2012), including probabilistic factors capturing genotype correlations as well as genotype–expression associations given by the eQTLs. A graphical representation of DSM is provided in Supplemental Figure S6.

We denote the two haplotypes of $\mathbf{x}$ as $\mathbf{h}^m$ (maternal) and $\mathbf{h}^p$ (paternal). DSM models the prior over each haplotype using the Li and Stephens' HMM (Li and Stephens 2003; Rubinacci et al. 2021) with respect to a given reference panel. The HMM includes a chain of hidden states $\mathbf{z}^m$ and $\mathbf{z}^p$ (for $\mathbf{h}^m$ and $\mathbf{h}^p$, respectively), which indicate, at each position, the index of reference panel haplotype from which the observed genotype is copied with a small probability of error; this copying distribution is captured by an emission factor, $\varepsilon(h_s|z_s)$ss, which relates the probability of the observed genotype at position s (h_s) conditioned on the indexed reference panel haplotype at that position (z_s). Each adjacent pair of hidden states are related through a transition factor, $\tau(z_s|z_{s-1})$, which defines the probability of crossover between the haplotypes at positions $s-1$ and s based on position-specific recombination rates (Loh et al. 2016). In addition to these two groups of factors that make up the pair of HMMs, DSM introduces a QTL factor ϕ_s for each eQTL s, capturing the dependence between the eQTL genotype and the expression levels of one or more genes, as defined in Equation 3.

The full joint probability distribution (conditional and unnormalized) can be written as

$$\begin{aligned}\tilde{p}_{\mathrm{DSM}}(\mathbf{x} = (\mathbf{h}^m, \mathbf{h}^p), \mathbf{z}^m, \mathbf{z}^p|\mathbf{e}) \\ = p_{\mathrm{HMM}}(\mathbf{h}^m, \mathbf{z}^m) \cdot p_{\mathrm{HMM}}(\mathbf{h}^p, \mathbf{z}^p) \cdot \tilde{p}_{\mathrm{QTL}}(\mathbf{x}|\mathbf{e}),\end{aligned} \tag{4}$$

where

$$p_{\mathrm{HMM}}(\mathbf{h}^m, \mathbf{z}^m) = \pi(z_1^m) \cdot \epsilon(h_1^m|z_1^m) \prod_{s=2}^{V} \tau(z_s^m|z_{s-1}^m) \cdot \epsilon(h_s^m|z_s^m), \tag{5}$$

$$\tilde{p}_{\mathrm{QTL}}(\mathbf{x}|\mathbf{e}) = \prod_{s=1}^{V} \phi_s(\mathbf{x}_s, \mathbf{e}_{Q(s)}), \tag{6}$$

and $p_{\mathrm{HMM}}(\mathbf{h}^p, \mathbf{z}^p)$ defined analogously to $p_{\mathrm{HMM}}(\mathbf{h}^m, \mathbf{z}^m)$. π represents the prior distribution over the first hidden state, typically a flat (uniform) prior.

To apply the model, we wish to compute the marginal distribution over $\mathbf{x}$ to use for the match score:

$$\tilde{p}_{\mathrm{DSM}}(\mathbf{x}|\mathbf{e}) = \sum_{\mathbf{z}^m,\mathbf{z}^p} \tilde{p}_{\mathrm{DSM}}(\mathbf{x}, \mathbf{z}^m, \mathbf{z}^p|\mathbf{e}). \tag{7}$$

We observe that this integration can be computed efficiently using an extension of the forward–backward algorithm commonly used for inference over HMMs (Browning and Browning 2007; Collins 2013; Rubinacci et al. 2021). This is achieved by considering each ϕ_s as a noisy observation of the corresponding genotype variables h_s^m and h_s^p and by incorporating them into the sequential belief updates along with the emission factors in the original algorithm. We provide the details of this algorithm in Supplemental Note S3.

Haplotype approximation

Given the quadratic state space at each position when jointly considering $\mathbf{z}^m$ and $\mathbf{z}^p$, the aforementioned inference procedure can be computationally expensive given a large reference panel. Using a large panel is necessary to accurately model the linkage disequilibrium patterns. Thus, we introduce the following modification of the QTL factors:

$$\phi_s(\mathbf{x}_s, \mathbf{e}_{Q(s)}) = \phi_s(h_s^m, \mathbf{e}_{Q(s)}) \cdot \phi_s(h_s^p, \mathbf{e}_{Q(s)}). \tag{8}$$

Intuitively, this approximates the predictive signal with a haplotype-specific effect that is equally applied to both haplotypes of the genotype profile. We then define

$$\phi_s(h_s^m, \mathbf{e}_{Q(s)}) = h_s^m \cdot \sigma(\alpha_s + \boldsymbol{\beta}_s^T\mathbf{e}_{Q(s)}) + (1 - h_s^m) \cdot (1 - \sigma(\alpha_s + \boldsymbol{\beta}_s^T\mathbf{e}_{Q(s)})), \tag{9}$$

and analogously for $\phi_s(h_s^p, \mathbf{e}_{Q(s)})$, where σ denotes the sigmoid function. The learnable parameters α_s and $\boldsymbol{\beta}_s$ are shared between the two haplotypes. The consequence of this modification is that now we have the factorization

$$\begin{aligned}\tilde{p}_{\mathrm{QTL}}(\mathbf{x}|\mathbf{e}) &= \left(\prod_{s=1}^{V} \phi_s(h_s^m, \mathbf{e}_{Q(s)})\right) \cdot \left(\prod_{s=1}^{V} \phi_s(h_s^p, \mathbf{e}_{Q(s)})\right) \\ &= \tilde{p}_{\mathrm{QTL}}(\mathbf{h}^m|\mathbf{e}) \cdot \tilde{p}_{\mathrm{QTL}}(\mathbf{h}^p|\mathbf{e}).\end{aligned} \tag{10}$$

This fully decouples the two haplotypes and allows us to express

$$\tilde{p}_{\mathrm{DSM}}(\mathbf{x}|\mathbf{e}) = \tilde{p}_{\mathrm{DSM}}(\mathbf{h}^m|\mathbf{e}) \cdot \tilde{p}_{\mathrm{DSM}}(\mathbf{h}^p|\mathbf{e}), \tag{11}$$

where each term is computed using the aforementioned forward–backward algorithm at the haplotype level. This reduces the overall runtime and memory requirements from quadratic to linear in the number of reference haplotypes, enabling us to leverage larger reference panels. Similarly, we obtain a memory requirement linear in the number of eQTLs considered (see Supplemental Fig. S4).

Joint learning of QTL factors

DSM provides a way to compute the probability of observing genotypes $\mathbf{x}$ given an expression profile $\mathbf{e}$, in which the adjustments

informed by eQTLs are encoded by the ϕ factors. To ensure that the ϕ factors are calibrated with respect to redundant predictive signals and genotype correlations, we jointly learn all ϕ factors to directly maximize $p_{\text{DSM}}(\mathbf{x}'^{(i)}|\mathbf{e}'^{(i)})$ across all matching pairs $(\mathbf{x}'^{(i)}, \mathbf{e}'^{(i)})$ in the training data set. This is in contrast to splitting the learning procedure to first separately training ϕ_s for each s to predict individual eQTLs and then using these predictions to compute a sequence-wide matching score, an approach followed by all of the existing works to our knowledge (Schadt et al. 2012; Harmanci and Gerstein 2016). Neglecting the dependence among different eQTLs while combining their predictions can lead to miscalibrated scores with poor predictive performance (see Supplemental Fig. S5).

Match score function

For evaluating the linking performance of DSM, we use the following match score:

$$M(\mathbf{x}, \mathbf{e}) = \tilde{p}_{\text{DSM}}(\mathbf{x}|\mathbf{e}) / p_{\text{HMM}}(\mathbf{x}). \quad (12)$$

The rationale for normalizing the output of DSM by $p_{\text{HMM}}(\mathbf{x})$ is as follows. Because we take a particular expression profile $\mathbf{e}$ and match it across candidate $\mathbf{x}$'s, $p(\mathbf{e}|\mathbf{x})$ is our desired choice for the score so as not to bias the match toward $\mathbf{x}$'s that are more likely just based on the prior. Note that $p(\mathbf{e}|\mathbf{x}) = p_{\text{DSM}}(\mathbf{x}|\mathbf{e})p(\mathbf{e}) / p_{\text{HMM}}(\mathbf{x})$, and because $p(\mathbf{e})$ and the normalization factor for $\tilde{p}_{\text{DSM}}(\mathbf{x}|\mathbf{e})$ are constant given a fixed $\mathbf{e}$, this is equivalent to matching using the above score M.

Implementation details

We implemented the forward–backward inference algorithm for DSMs in PyTorch (Collins 2013), leveraging the automatic differentiation features and the Adam optimizer for parameter learning. We set our learning rate to 0.025 and number of epochs to 50 based on cross-validation, which was performed on a sample window of 750 eQTLs and by holding out a subset of 58 GTEx individuals (out of 588) for validation. Selected hyperparameters were used on the full GTEx data set for all our experiments. For parallelization, we trained a separate DSM for each genomic window of 750 eQTLs with 75 GB RAM and 4 CPUs for a runtime of ~8 h. The evaluation of runtime and memory scaling with respect to window size and reference panel size is provided (see Supplemental Fig. S4).

For the baseline methods (GNB and EBL), we implemented the pruning of eQTLs by greedily selecting the most significant eQTL (as reported in the training data set) and then removing any other eQTLs that are correlated with it. The choice of correlation threshold was a hyperparameter we optimized, and we found that removing eQTLs with correlation greater than 0.1 to the significant eQTL consistently led to best linking accuracy on FUSION individuals. We thus adopted this threshold for comparison with our approach. Correlation was calculated as the Pearson correlation coefficient of observed genotypes between each pair of eQTLs in the training set.

For EBL, we additionally optimized the eQTL–eGene correlation threshold (for determining the inclusion of eQTLs) and the extremity threshold (for assinging extreme gene expression levels to the corresponding homozygous genotypes) over the range of values considered in the software provided by the original publication. Notably, the original approach evaluates the linking performance for fixed thresholds and does not require a separate training phase. Nevertheless, we optimized these parameters on our training set to compare with other approaches that do leverage such training data.

Data sets

We obtained our data sets through the NCBI dbGaP (Tryka et al. 2014) and the European Genome-Phenome Archive (EGA; https://ega-archive.org) (Lappalainen et al. 2015). From dbGaP, we downloaded the GTEx v8 muscle tissue expression (phe000037.v1) and genotype (phg001219.v1) data sets (The GTEx Consortium 2015) and the FUSION expression (phe000033.v1) and imputed genotype (phg001194.v1) data sets (Ghosh et al. 2000). We downloaded the HRC reference panel from the EGA (EGAS00001001710) (The Haplotype Reference Consortium 2016). Genotype samples in our data sets that were not already phased were phased with the Michigan Imputation Server (Das et al. 2016) using the Eagle2 software (Loh et al. 2016). Gene expression profiles are normalized with PEER factor normalization using the default parameter setting (Stegle et al. 2012).

We included in our models all eQTL associations reported in the GTEx data set that overlapped with the genotype data in FUSION, which consisted of following: 47,322 eQTLs and 735 eGenes for Chromosome 19; 21,685 eQTLs and 254 eGenes for Chromosome 20; 11,740 eQTLs and 118 eGenes for Chromosome 21; 19,241 eQTLs and 264 eGenes for Chromosome 22; and 99,988 eQTLs and 1011 eGenes for all four chromosomes combined.

Software availability

Our Python implementations of DSM training and linking algorithms and example data formats and scripts are provided as Supplemental Code and are also available at GitHub (https://github.com/shuvom-s/DSM).

Competing interest statement

The authors declare no competing interests.

Acknowledgments

This work is accepted for an oral presentation at RECOMB 2023. This work is supported by the Hertz Fellowship and National Science Foundation Graduate Research Fellowship under grant number 2141064 (to S.S.), National Institutes of Health (NIH) R01 HG010959 (to B.B.), and NIH DP5 OD029574 and RM1 HG011558 (to H.C.).

Author contributions: S.S., D.F., and H.C. conceived the project. S.S. and H.C. designed the experiments and developed the method. S.S. and D.F. implemented the software and performed the experiments. B.B. and H.C. supervised the project. All authors analyzed the data and wrote the manuscript.

References

Almadhoun N, Ayday E, Ulusoy Ö. 2020. Differential privacy under dependent tuples: the case of genomic privacy. *Bioinformatics* **36:** 1696–1703. doi:10.1093/bioinformatics/btz837

Athar A, Füllgrabe A, George N, Iqbal H, Huerta L, Ali A, Snow C, Fonseca NA, Petryszak R, Papatheodorou I, et al. 2019. ArrayExpress update: from bulk to single-cell expression data. *Nucleic Acids Res* **47:** D711–D715. doi:10.1093/nar/gky964

Backes M, Berrang P, Bieg M, Eils R, Herrmann C, Humbert M, Lehmann I. 2017. Identifying personal DNA methylation profiles by genotype inference. In *Proceedings of IEEE Symposium on Security and Privacy (S&P) 2017*, San Jose, CA, pp. 957–976. IEEE.

Barbeira A, Shah KP, Torres JM, Wheeler HE, Torstenson ES, Edwards T, Garcia T, Bell GI, Nicolae D, Cox NJ, et al. 2018. Exploring the phenotypic consequences of tissue specific gene expression variation inferred

from GWAS summary statistics. *Nat Commun* **9:** 1825. doi:10.1038/s41467-018-03621-1

Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, et al. 2010. NCBI GEO: archive for functional genomics data sets—10 years on. *Nucleic Acids Res* **39(Suppl_1):** D1005–D1010. doi:10.1093/nar/gkq1184

Bonomi L, Huang Y, Ohno-Machado L. 2020. Privacy challenges and research opportunities for genomic data sharing. *Nat Genet* **52:** 646–654. doi:10.1038/s41588-020-0651-0

Browning SR, Browning BL. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* **81:** 1084–1097. doi:10.1086/521987

Browning BL, Tian X, Zhou Y, Browning SR. 2021. Fast two-stage phasing of large-scale sequence data. *Am J Hum Genet* **108:** 1880–1890. doi:10.1016/j.ajhg.2021.08.005

Chen J, Wang WH, Shi X. 2021. Differential privacy protection against membership inference attack on machine learning for genomic data. In *Proceedings of the Pacific Symposium on Biocomputing (PSB) 2021*, Kohala Coast, HI, pp. 26–37. World Scientific.

Cho H, Froelicher D, Chen J, Edupalli M, Pyrgelis A, Troncoso-Pastoriza JR, Hubaux J-P, Berger B. 2022. Secure and federated genome-wide association studies for Biobank-scale datasets. bioRxiv doi:10.1101/2022.11.30.518537

Clayton EW, Evans BJ, Hazel JW, Rothstein MA. 2019. The law of genetic privacy: applications, implications, and limitations. *J Law Biosci* **6:** 1–36. doi:10.1093/jlb/lsz007

Collins M. 2013. *The forward-backward algorithm*. Columbia University, New York.

Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew EY, Levy S, McGue M, et al. 2016. Next-generation genotype imputation service and methods. *Nat Genet* **48:** 1284–1287. doi:10.1038/ng.3656

Delaneau O, Zagury J-F, Robinson MR, Marchini JL, Dermitzakis ET. 2019. Accurate, scalable and integrative haplotype estimation. *Nat Commun* **10:** 5436. doi:10.1038/s41467-019-13225-y

Deznabi I, Mobayen M, Jafari N, Tastan O, Ayday E. 2018. An inference attack on genomic data using kinship, complex correlations, and phenotype information. *IEEE/ACM Trans Comput Biol Bioinform* **15:** 1333–1343. doi:10.1109/TCBB.2017.2709740

Dokmai N, Kockan C, Zhu K, Wang X, Sahinalp SC, Cho H. 2021. Privacy-preserving genotype imputation in a trusted execution environment. *Cell Syst* **12:** 983–993.e7. doi:10.1016/j.cels.2021.08.001

Emani PS, Gürsoy G, Miranker A, Gerstein MB. 2021. PLIGHT: a tool to assess privacy risk by inferring identifying characteristics from sparse, noisy genotypes. bioRxiv doi:10.1101/2021.07.18.452853

Erlich Y, Narayanan A. 2014. Routes for breaching and protecting genetic privacy. *Nat Rev Genet* **15:** 409–421. doi:10.1038/nrg3723

Erlich Y, Shor T, Pe'er I, Carmi S. 2018. Identity inference of genomic data using long-range familial searches. *Science* **362:** 690–694. doi:10.1126/science.aau4832

Flannick J, Mercader JM, Fuchsberger C, Udler MS, Mahajan A, Wessel J, Teslovich TM, Caulkins L, Koesterer R, Barajas-Olmos F, et al. 2019. Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature* **570:** 71–76. doi:10.1038/s41586-019-1231-2

Froelicher D, Troncoso-Pastoriza JR, Raisaro JL, Cuendet MA, Sousa JS, Cho H, Berger B, Fellay J, Hubaux J-P. 2021. Truly privacy-preserving federated analytics for precision medicine with multiparty homomorphic encryption. *Nat Commun* **12:** 5910. doi:10.1038/s41467-021-25972-y

Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, Eyler AE, Denny JC, Nicolae DL, Cox NJ, et al. 2015. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* **47:** 1091–1098. doi:10.1038/ng.3367

Gaunt TR, Shihab HA, Hemani G, Min JL, Woodward G, Lyttleton O, Zheng J, Duggirala A, McArdle WL, Ho K, et al. 2016. Systematic identification of genetic influences on methylation across the human life course. *Genome Biol* **17:** 61. doi:10.1186/s13059-016-0926-z

Ghosh S, Watanabe RM, Valle TT, Hauser ER, Magnuson VL, Langefeld CD, Ally DS, Mohlke KL, Silander K, Kohtamäki K, et al. 2000. The Finland–United States investigation of non–insulin-dependent diabetes mellitus genetics (fusion) study. I. An autosomal genome scan for genes that predispose to type 2 diabetes. *Am J Hum Genet* **67:** 1174–1185. doi:10.1016/S0002-9297(07)62948-6

The GTEx Consortium. 2015. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348:** 648–660. doi:10.1126/science.1262110

Gürsoy G, Emani P, Brannon CM, Jolanki OA, Harmanci A, Strattan JS, Cherry JM, Miranker AD, Gerstein M. 2020. Data sanitization to reduce private information leakage from functional genomics. *Cell* **183:** 905–917.e16. doi:10.1016/j.cell.2020.09.036

Gürsoy G, Lu N, Wagner S, Gerstein M. 2021. Recovering genotypes and phenotypes using allele-specific genes. *Genome Biol* **22:** 263. doi:10.1186/s13059-021-02477-x

Gürsoy G, Brannon CM, Ni E, Wagner S, Khanna A, Gerstein M. 2022a. Storing and analyzing a genome on a blockchain. *Genome Biol* **23:** 134. doi:10.1186/s13059-022-02699-7

Gürsoy G, Li T, Liu S, Ni E, Brannon CM, Gerstein MB. 2022b. Functional genomics data: privacy risk assessment and technological mitigation. *Nat Rev Genet* **23:** 245–258. doi:10.1038/s41576-021-00428-7

Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BW, Jansen R, De Geus EJ, Boomsma DI, Wright FA, et al. 2016. Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* **48:** 245–252. doi:10.1038/ng.3506

Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. 2013. Identifying personal genomes by surname inference. *Science* **339:** 321–324. doi:10.1126/science.1229566

The Haplotype Reference Consortium. 2016. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48:** 1279–1283. doi:10.1038/ng.3643

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

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

He Z, Yu J, Li J, Han Q, Luo G, Li Y. 2020. Inference attacks and controls on genotypes and phenotypes for individual genomic data. *IEEE/ACM Trans Comput Biol Bioinf* **17:** 930–937. doi:10.1109/TCBB.2018.2810180

Hill AC, Guo C, Litkowski EM, Manichaikul AW, Yu B, Konigsberg IR, Gorbet BA, Lange LA, Pratte KA, Kechris KJ, et al. 2023. Large scale proteomic studies create novel privacy considerations. *Sci Rep* **13:** 9254. doi:10.1038/s41598-023-34866-6

Lambert C, Fernandes M, Decouchant J, Esteves-Verissimo P. 2018. MaskAl: Privacy preserving masked reads alignment using Intel SGX. In *2018 IEEE 37th Symposium on Reliable Distributed Systems (SRDS)*, Salvador, Brazil, pp. 113–122. IEEE.

Lappalainen I, Almeida-King J, Kumanduri V, Senf A, Spalding JD, Saunders G, Kandasamy J, Caccamo M, Leinonen R, Vaughan B, et al. 2015. The European genome-phenome archive of human data consented for biomedical research. *Nat Genet* **47:** 692–695. doi:10.1038/ng.3312

Li N, Stephens M. 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165:** 2213–2233. doi:10.1093/genetics/165.4.2213

Loh P-R, Danecek P, Palamara PF, Fuchsberger C, A Reshef Y, K Finucane H, Schoenherr S, Forer L, McCarthy S, Abecasis GR, et al. 2016. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet* **48:** 1443–1448. doi:10.1038/ng.3679

Naveed M, Ayday E, Clayton EW, Fellay J, Gunter CA, Hubaux J-P, Malin BA, Wang X. 2015. Privacy in the genomic era. *ACM Comput Surv* **48:** 6. doi:10.1145/2767007

Rubinacci S, Ribeiro DM, Hofmeister RJ, Delaneau O. 2021. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nat Genet* **53:** 120–126. doi:10.1038/s41588-020-00756-0

Samani SS, Huang Z, Ayday E, Elliot M, Fellay J, Hubaux J-P, Kutalik Z. 2015. Quantifying genomic privacy via inference attack with high-order SNV correlations. In *2015 IEEE Security and Privacy Workshops (SPW)*, San Jose, CA, pp. 32–40. IEEE.

Schadt E, Woo S, Hao K. 2012. Bayesian method to predict individual SNP genotypes from gene expression data. *Nat Genet* **44:** 603–608. doi:10.1038/ng.2248

Stegle O, Parts L, Piipari M, Winn J, Durbin R. 2012. Using probabilistic estimation of expression residuals (peer) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc* **7:** 500–507. doi:10.1038/nprot.2011.457

Sutton C, McCallum A. 2012. An introduction to conditional random fields. *Found Trends Mach Learn* **4:** 267–373. doi:10.1561/2200000013

Sweeney L. 2000. Simple demographics often identify people uniquely. *Health (San Francisco)* **671:** 1–34.

Tramèr F, Huang Z, Hubaux J-P, Ayday E. 2015. Differential privacy with bounded priors: reconciling utility and privacy in genome-wide association studies. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS)*, Denver, CO, pp. 1286–1297.

Tryka KA, Hao L, Sturcke A, Jin Y, Wang ZY, Ziyabari L, Lee M, Popova N, Sharopova N, Kimura M, et al. 2014. NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res* **42:** D975–D979. doi:10.1093/nar/gkt1211

Wan Z, Hazel JW, Clayton EW, Vorobeychik Y, Kantarcioglu M, Malin BA. 2022. Sociotechnical safeguards for genomic data privacy. *Nat Rev Genet* **23:** 429–445. doi:10.1038/s41576-022-00455-y

Ye F, Cho H, El Rouayheb S. 2022. Mechanisms for hiding sensitive genotypes with information-theoretic privacy. *IEEE Trans Inf Theory* **68:** 4090–4105. doi:10.1109/TIT.2022.3156276

Yilmaz E, Ayday E, Ji T, Li P. 2020. Preserving genomic privacy via selective sharing. In *Proceedings of the 19th Workshop on Privacy in the Electronic Society*, pp. 163–179. Association for Computing Machinery, New York.

Zhang C, Bonomi L. 2022. Mitigating membership inference in deep learning applications with high dimensional genomic data. In *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)*, Rochester, MN, pp. 01–03. IEEE.

Received January 14, 2023; accepted in revised form April 19, 2023.