Privacy-Preserving Collaborative Genomic Research: A Real-Life Deployment and Vision

Zahra Rahmani*
Nahal Shahini*
zxr81@case.edu
nxs814@case.edu
Case Western Reserve University
Cleveland, Ohio, USA

Nadav Gat†
Zebin Yun†
nadavgat@mail.tau.ac.il
zebinyun@mail.tau.ac.il
Tel Aviv University
Tel Aviv, Israel

Yuzhou Jiang
yxj466@case.edu
Case Western Reserve University
Cleveland, Ohio, USA

Ofir Farchy
ofir@lynx.md
Lynx.MD
Palo Alto, California, USA

Yaniv Harel
yaniv10@tauex.tau.ac.il
Tel Aviv University
Tel Aviv, Israel

Vipin Chaudhary
vxc204@case.edu
Case Western Reserve University
Cleveland, Ohio, USA

Erman Ayday‡
exa208@case.edu
Case Western Reserve University
Cleveland, Ohio, USA

Mahmood Sharif‡
mahmoods@tauex.tau.ac.il
Tel Aviv University
Tel Aviv, Israel

Abstract

The data revolution holds a significant promise for the health sector. Vast amounts of data collected and measured from individuals will be transformed into knowledge, AI models, predictive systems, and digital best practices. One area of health that stands to benefit greatly from this advancement is the genomic domain. The advancement of AI, machine learning, and data science has opened new opportunities for genomic research, promising breakthroughs in personalized medicine. However, the increasing awareness of privacy and cyber security necessitates robust solutions to protect sensitive data in collaborative research. This paper presents a practical deployment of a privacy-preserving framework for genomic research, developed in collaboration with Lynx.MD, a platform designed for secure health data collaboration. The framework addresses critical cyber security and privacy challenges, enabling the privacy-preserving sharing and analysis of genomic data while mitigating risks associated with data breaches. By integrating advanced privacy-preserving algorithms, the solution ensures the protection of individual privacy without compromising data utility. A unique feature of the system is its ability to balance the trade-offs between data sharing and privacy, providing stakeholders with tools to quantify privacy risks and make informed decisions. The implementation of the framework within Lynx.MD involves encoding genomic data into binary formats and applying noise through controlled perturbation techniques. This approach preserves essential statistical properties of the data, facilitating effective research and analysis. Additionally, the system incorporates real-time data monitoring and advanced visualization tools, enhancing user experience and decision-making capabilities. The paper highlights the need for tailored privacy attacks and defenses specific to genomic data, given its unique characteristics compared to other data types. By addressing these challenges, the proposed solution aims to foster global collaboration in genomic research, ultimately contributing to significant advancements in personalized medicine and public health.

CCS Concepts

• **Security and privacy → Privacy-preserving protocols**; • **Usability in security and privacy**;

Keywords

[Data sharing; Genomic privacy; Usable privacy]

ACM Reference Format:
Zahra Rahmani, Nahal Shahini, Nadav Gat, Zebin Yun, Yuzhou Jiang, Ofir Farchy, Yaniv Harel, Vipin Chaudhary, Erman Ayday, and Mahmood Sharif. 2024. Privacy-Preserving Collaborative Genomic Research: A Real-Life Deployment and Vision. In *Proceedings of the 2024 Workshop on Cybersecurity in Healthcare (HealthSec '24), October 14–18, 2024, Salt Lake City, UT, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3689942.3694747

1 Introduction

The world has an incremental progress in collecting genomic data over the years. The new generation of capabilities including AI, machine learning, and data science, provide a new opportunity to investigate and run research that may identify new variants, connect symptoms to root causes, and develop new (personalized)

*Both authors contributed equally to this research.
†Both authors contributed equally to this research.
‡Corresponding authors.

HealthSec '24, October 14–18, 2024, Salt Lake City, UT, USA

ACM ISBN 979-8-4007-1238-8/24/10
https://doi.org/10.1145/3689942.3694747

medicines. Every healthcare institution, a clinic, a health research center, and a medical organization, holds unique information that can highly contribute to the global research of specific cases and problems. While this is clear, at the same time, humanity increases the awareness and the perceived importance of privacy and cyber security. Making the maximum effort to protect the patients' information and personal data of individuals that participate in clinical trials is of high importance. An important development in defining the responsibilities of these procedures' organizers has helped protect private data against unintentional leakage or malicious attacks. The outcome of this progress between other reasons creates a situation that most of this valuable data stays at the organizations that gathered the information rather than data sharing and collaborative research across institutions.

In this work, we provide our vision for privacy-preserving collaborative genomic research and showcase our first attempt for practical, a real-life deployment. In order to make this work practical and connected to real use-cases we work with an industry collaborator, Lynx.MD, and use the existing genomic data storage platform of this partner. Lynx.MD is an American Israeli start-up that has developed a platform for health data collaboration between healthcare institutions, pharma companies, and research foundations. The practical use-cases and the real data are essential for the ability to deeply study the research goals. The proposed framework mainly addresses the cyber security and privacy issues that can better protect the data and reduce the damage in case of any data breach. It also allows researchers (users of Lynx.MD, who store their datasets in the platform) share their datasets among each other (e.g., for collaborative research) in a privacy-preserving way and share their AI models as black-box.

We also propose including a component in the system that demonstrates the trade-off between information sharing and the associated privacy costs. Currently, there are no interpretable tools available to measure privacy loss, resulting in decisions about data sharing being made without full awareness. Individuals responsible for these decisions are not always privacy and security experts, making it essential to develop methodologies that clearly present privacy penalties and provide the necessary context for informed decision-making.

Overall, this work addresses one of the most innovative domains that requires solving key challenges in order to make a dramatic change on the global ability to leverage distributed health data. Doing so will contribute to creating more collective data repositories that may enable the use of AI to find solutions, make diagnoses, and develop medicines for the most prioritized world diseases. To achieve our vision, we have already developed and integrated privacy-preserving algorithms with the Lynx.MD platform, created interpretable privacy risk tools, and deployed these mechanisms in high-performance computing environments. These steps ensure robust data protection and facilitate collaborative research. Moving forward, we will focus on evaluating and validating these tools with real-world data, fostering collaboration and training, and continuously enhancing cybersecurity measures. This roadmap aims to ensure robust data protection, facilitate collaborative research, and enable the practical use of AI and machine learning in genomic research while addressing privacy and security concerns. Note that although this work focuses on privacy, cybersecurity aspects are an integral part of protecting assets and data confidentiality. AI models are part of the protected landscape, along with the technologies used to better defend processes and data flow [20, 26, 29]. Future work will explore this aspect in greater depth.

2 Related Works

We discuss related work from four aspects: privacy risks for genomic data, privacy-preserving solutions for genomic research, and machine learning (ML) and AI privacy.

2.1 Privacy Risks for Genomic Data

Genomic privacy has recently been explored by many researchers [4, 5, 23, 51, 59]. Several works have shown that anonymization does not effectively protect the privacy of genomic data [27, 28, 30, 39, 42, 45, 58]. Previous research have demonstrated that the identity of a participant of a genomic study can be revealed by using a second sample, that is, part of the DNA information from the individual and the results of the clinical study [19, 31, 33, 54, 57, 61, 63, 67]. Furthermore, several studies have examined phenotype prediction from genomic data, as a means of tracing identity [3, 18, 32, 41, 43, 44, 46, 52, 62, 69]. This line of research highlights the vulnerabilities in genomic data sharing, particularly when datasets are linked with other publicly available data sources. The predictive power of genomic data can be exploited to infer sensitive information about individuals, such as their susceptibility to certain diseases or traits like eye and hair color, which can then be used for discriminatory purposes.

Recent studies have also highlighted the potential for privacy breaches through familial search techniques, where an adversary can identify individuals by matching their genetic markers with those of their relatives [28]. This method leverages the genetic similarity among family members to breach privacy, making it a significant concern for genomic data sharing practices. The growing availability of public genetic databases intensifies this risk, as these repositories provide a rich source of genetic information that can be leveraged to cross-reference and identify individuals from anonymized datasets.

2.2 Privacy-Preserving Solutions for Genomic Research

Most proposed solutions to utilize or share genomic data are either based on obfuscation or the use of cyptographic techniques. Some researchers have proposed using differential privacy (DP) concept to mitigate membership inference attacks when releasing summary statistics, such as minor allele frequencies and chi-square values [24, 38, 66]. DP provides a mathematical framework that ensures the addition or removal of a single database item does not significantly affect the outcome of any analysis, thereby preserving the privacy of individuals in the dataset.

A significant subset of cryptographic solutions focuses on private pattern-matching and comparison of genomic sequences [11, 21, 34, 50, 60]. For privacy-preserving clinical genomics, Baldi *et al.* proposed private set intersection-based techniques [10]. These techniques allow researchers to identify common genetic variants across

datasets without revealing individual data points. Similarly, partially homomorphic encryption has been proposed for the privacy-preserving use of genomic data in clinical settings, enabling computations to be performed on encrypted data without revealing the underlying information [6–9]. Kantarcioglu *et al.* proposed homomorphic encryption-based techniques for privacy-preserving genomic research, which allow secure computations on genetic data without revealing the data itself [40]. Wang *et al.* proposed private edit distance protocols to find similar patients across multiple hospitals, enhancing collaborative research without compromising patient privacy [64]. These cryptographic methods offer robust privacy guarantees and are particularly suited for applications where sensitive genomic data must be processed securely.

Despite these advancements, cryptographic solutions face challenges related to interoperability and practicality. Different genomic analysis problems require distinct cryptosystems, leading to security and efficiency issues when integrating multiple systems. The computational intensity of some techniques also poses challenges for large-scale studies.

Future research should prioritize developing unified frameworks that offer comprehensive privacy-preserving solutions while maintaining high data utility and efficiency. This includes designing versatile cryptographic systems that can handle diverse genomic analysis tasks and ensuring scalability to manage the increasing volume of genomic data generated by modern sequencing technologies. Advancing these areas will enable genomic research to progress without compromising data privacy and security.

2.3 ML/AI Privacy

A wide range of privacy attacks have been proposed against ML models in recent years, demonstrating various ways in which ML may leak sensitive data during or after training [15, 25, 56, 68]. For example, membership inference attacks enable adversaries with access to a trained model to determine whether a record of data was used in training [56]. As another example, model inversion allows adversaries to reconstruct training samples from trained models [25]. Other attacks include, but are not limited to unintended memorization [15] and input and label reconstruction in federated learning [68].

ML privacy attacks can serve as a means to empirically quantify leakage of private data under various settings. Still, to our knowledge, these attacks have primarily been explored and used by ML privacy academics. In our vision (Sec. 3), the attack outcomes would be made available to the stakeholders on the Lynx.MD platform to aid them in decision making when they need to decide whether to release a model trained on the platform. We also note that ML privacy attacks have been mainly studied in the vision and text domains, with little effort exploring their effectiveness on genomic and health data. Accordingly, it remains unclear whether established attacks can reliably assess privacy leakage in ML in these domains, suggesting that new attacks tailored for genomic and health data may need to be developed.

Researchers have also proposed various methods to enhance ML privacy [2, 12, 49, 53]. While some of these methods do not provably guarantee privacy but have been demonstrated empirically effective [49], other methods also carry provable guarantees [2, 53]. Notably, the differentially private stochastic gradient descent algorithm [2] enables training ML models via stochastic gradient descent while satisfying differential privacy guarantees, thus, in a sense, providing plausible deniability about whether certain records were used in training. In our vision, these defenses would be provided as a service to users of the Lynx.MD platform, thus helping protect data privacy when training models (Sec. 3).

3 Our Vision

To address the critical challenges associated with genomic data privacy, we developed a comprehensive privacy-preserving solution for integration with the Lynx.MD platform. This solution leverages advanced privacy-preserving solutions to ensure robust privacy protections while maintaining the utility of shared genomic data.

In our broader vision, Lynx.MD serves as a sandbox platform where researchers can securely upload and share their genomic datasets. Our privacy preserving algorithms are seamlessly integrated into Lynx.MD, providing users with tools to analyze and share data without exposing sensitive information. Additionally, a user-friendly interface and sophisticated visualization tools will be developed to allow users to easily manage their datasets and understand the privacy and utility levels of shared data. These tools are crucial for making informed decisions about data sharing and analysis.

3.1 Data Sharing Between Lynx.MD Users

We employed a state-of-the-art privacy-preserving genomic dataset sharing algorithm [36], which operates in two critical stages: data perturbation and utility restoration. Initially, the genomic SNP data, with values 0, 1, and 2, is encoded into binary space where each SNP value is transformed: 0 to 00, 1 to 01, and 2 to 11, resulting in a binary matrix of size n×2m. Following this, a noise matrix of the same dimensions is generated using a binary-valued XOR mechanism [35], which adds noise to the encoded data to obscure the original genomic sequences. The noise matrix is created with Bernoulli-distributed values, accounting for both row (individual) and column (SNP)-wise correlations as described in the original work [35]. However, due to the computational complexity of the XOR mechanism in large genomic datasets, we have developed an enhanced version of noise sampling called efficient binary noise generation [37]. This enhancement allows for the efficient generation of the noise matrix, which is then XORed with the encoded data, producing a noisy binary matrix. The noisy matrix is subsequently decoded back into the original SNP space before sharing, though this occurs prior to any utility-enhancing post-processing.

To further preserve data utility, particularly for genome-wide association studies (GWAS), we employ a utility restoration stage that adjusts the minor allele frequencies (MAFs) [16] in the noisy dataset to better align with publicly available MAF values. This post-processing method uses an optimal transport approach, specifically the earth mover's distance [55], to determine the minimal number of alleles to flip to achieve the desired alignment. The process is as follows:

(1) Calculate the MAF value $\tilde{\mathcal{M}}_j$ of each SNP j in the noisy (binarized) dataset $\tilde{D}^b$.

(2) Compute the transition from $\tilde{\mathcal{M}}_j$ to the reference MAF $\tilde{\mathcal{M}}_j^r$ using the earth mover's distance, which estimates the percentage of alleles to be flipped.
(3) Determine the exact number of alleles to flip by applying the floor function to the percentage multiplied by the total number of alleles.
(4) Randomly select and flip the necessary alleles.

This post-processing effectively enhances the reproducibility of GWAS and minimizes the point error, improving the overall utility of the shared dataset. The robustness of our approach is further demonstrated by its ability to maintain the privacy of individuals' genomic data while allowing researchers to replicate significant findings reliably.

This two-stage privacy-preserving scheme is essential for maintaining the confidentiality of genomic data while preserving its utility for research purposes. Our implementation within a high-performance computing (HPC) environment ensures that the process is scalable and capable of handling large datasets. The Lynx.MD platform, which serves as a controlled environment for data sharing, incorporates this mechanism to allow researchers to securely upload their datasets. The platform's design also supports collaborative research by providing secure access to shared datasets and tools necessary for joint studies. Once datasets are uploaded, the Lynx.MD platform applies the privacy-preserving scheme, ensuring the data remains protected while still being usable for advancing personalized medicine and other research initiatives. The use of a privacy budget ϵ allows us to balance the privacy-utility tradeoff effectively, with an optimal range of ϵ between 1 and 1000, equivalent to a privacy budget of less than approximately 0.2 per SNP [37]. This tradeoff is crucial for maintaining data utility while safeguarding individual privacy.

Additionally, the verifier, who may be a reviewer in a peer review process or another researcher seeking to validate the results, can reproduce the researcher's experiments using a sanitized version of the dataset. This process includes calculating the SNP retention rate, a metric that indicates the percentage of SNPs that remain statistically significant in reproduced results compared to those reported by the researcher. The verifier can also use public information during this process, potentially applying a relaxed p-value threshold to assess the SNP retention rate. By comparing the retention rate to a theoretical ideal or expected rate, the verifier can assess the reliability of the findings. If the difference between the actual and expected rates falls within a specified threshold, the findings are deemed reliable. Otherwise, further investigation may be initiated, or additional detailed information may be requested, pending Institutional Review Board (IRB) approval. The public availability of MAF statistics, which is allowed according to NIH guidelines, further strengthens the privacy guarantees of our method, as differential privacy's immunity to post-processing ensures that these statistics do not compromise the overall privacy of the data [22].

3.2 Privacy Assurance in ML Models on the Lynx.MD Platform

A fundamental premise of the Lynx.MD platform is that training machine learning (ML) models on the data available through the platform enables technologists and scientists to develop innovative technologies and derive valuable scientific insights. However, as previously discussed (Sec. 2.3), releasing these models carries significant privacy risks. Adversaries may leverage access to the models to infer sensitive information about individuals' genomic or health data. Thus, it is imperative to ensure that the models leak minimal to no information about the training data before their release from the platform. To this end, our vision includes providing stakeholders (both data owners and users) with an automated evaluation system that assesses the extent to which models leak information about their training data. Additionally, stakeholders will receive recommendations for potential defenses that can be incorporated during training to enhance the protection of the training data.

Various privacy attacks have been proposed and thoroughly evaluated against ML models, as previously mentioned (Sec. 2.3). However, genomic datasets exhibit significant differences compared to the image or text datasets commonly used in past evaluations [14, 56]. In return, these differences can impact the effectiveness of attacks. First, genomic datasets typically have relatively small sample sizes, as collecting genomic data remains more expensive than collecting text or image data, which can be easily gathered from the internet, albeit not always labeled. Some attacks require auxiliary datasets to train surrogate models that are later used to infer membership [14]. These may be ineffective against genomic models due to data scarcity. Second, while there are millions of genomic features—such as Single Nucleotide Polymorphisms (SNPs) [13] that can be ingested by models, the number of useful features in a given dataset is often substantially smaller, with each feature admitting a limited set of values. In contrast, image and text datasets contain significantly more dimensions (e.g., 3,072–268,203 dimensions for standard image datasets), each admitting 256 different values [56]. The lower dimensionality of genomic data may decrease overfitting, rendering certain attacks such as membership inference more challenging [65]. Furthermore, many genomic prediction models (e.g., Ordinary Least Squares, Classical Ridge Regression, Linear regression, Classical Elastic Net and Bayesian Ridge regression [47]) are inherently more transparent and less prone to overfitting, making them robust choices for genomic data analysis. These attributes ensure that the models provide reliable and interpretable results, even when integrated into privacy-preserving frameworks. All in all, it may be necessary to develop novel attacks specifically tailored for genomic data, and health data in general, to reliably assess ML models' leakage. Indeed, as our evaluation of existing attacks showed limited success, we intend to plan to explore new attack directions in future work (e.g., by leveraging augmentations better-suited for genomic data [17]).

3.3 Communicating the Privacy Risk to Users

Once the risk of private information leakage due to data sharing or ML model release is assessed, it is necessary to communicate the risk to users to facilitate their decision making regarding whether to share data or models. To this end, among others, it is crucial to communicate to users the (1) potential implications of certain risks (e.g., membership inference); (2) theoretical guarantees of countermeasures applied (e.g., differential privacy mechanisms); (3) empirical assessment of data leakage (e.g., for potential success of membership inference); (4) assumptions behind the attacks for

which the risk is estimated (e.g., whether the adversary has auxiliary data); (5) potential tradeoffs between utility and privacy that can be attained with various countermeasures. We intend to devise a privacy dashboard that conveys such information to stakeholders in an usable manner. Primarily, for usability, it is essential to optimize accessibility and run time. For the former, we plan to rely on established literature offering ways to describe the sophisticated theoretical guarantees of certain defenses [48] and conduct user studies to find adequate means to convey metrics estimated by ML privacy attacks that would be accessible by stakeholders with a wide-range of backgrounds and expertise. For the latter, we intend to explore means to enable prompt assessment of necessary metrics (e.g., by proposing more efficient attacks).

4 Proof-of-Concept Implementation

We integrated the privacy-preserving genomic data sharing scheme into the Lynx.MD platform and rigorously assessed both the privacy and utility of the data using comprehensive metrics. Our analysis employed a variety of tests, including Average Point Error, Average Sample Error, and Mean and Variance Error, which together indicate how well the integrity and statistical properties of the original dataset are preserved post-transformation. The privacy level of the data was measured using the differential privacy parameter, ϵ, which provides strong privacy protection at different levels.

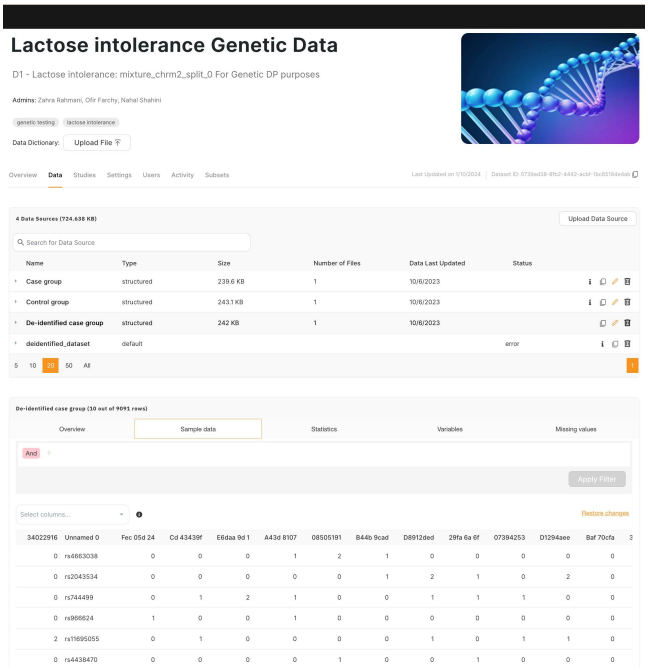

Figure 1: Lynx.MD Platform

4.1 Dataset

The dataset utilized for evaluating our privacy-preserving solution comprises a comprehensive collection of genomic data typical of what might be used in advanced medical research. This dataset features a broad spectrum of SNP variations, representing a diverse genetic background to ensure the generalizability of our results. Prior to the application of our privacy-preserving mechanisms, genomic data was encoded into a binary matrix format. This preparatory step was critical for facilitating the subsequent integration of our novel two-stage privacy-preserving algorithm, which applies controlled noise and utilizes publicly known statistics to enhance data quality.

We implemented and evaluated our proposed scheme on real-life genomic datasets from the OpenSNP project [1], which is a public platform that allows users to share their genetic data, typically derived from consumer genetic testing services. We selected three phenotypes for our study: lactose intolerance, hair color, and eye color. The **lactose intolerance** dataset includes 9,091 SNPs from 60 individuals with lactose intolerance. The **hair color** dataset contains 9,686 SNPs from 60 individuals with dark hair. The **eye color** dataset is larger, with 28,396 SNPs among 401 individuals with brown eyes. Additionally, a **handedness** dataset is included with 28,396 SNPs among 401 individuals.

We built a reference dataset for each phenotype, aligning the SNPs with the target dataset to be shared. These reference datasets were constructed from the remaining data in the OpenSNP project. This thorough dataset selection and preparation ensured that our evaluation was robust and reflective of real-world scenarios. Note that we included these datasets on the Lynx.MD platform as users of the platform and ran the algorithms on the platform using these datasets as shown in Figure 1.

4.2 Utility and Privacy Evaluation

Our results in [36] demonstrate that the proposed scheme consistently outperforms existing methods in maintaining data utility, regardless of the privacy budget. This strong performance highlights how effective our two-stage privacy-enhancing mechanism is. By adding controlled noise and using a unique post-processing technique, we achieve an excellent balance between data privacy and usability. This is essential for moving forward with collaborative genomic research and personalized medicine.

On the other hand, the differential privacy framework safeguards against inference attacks, such as membership inference attacks, by injecting noise and protecting SNP value distributions, thereby offering robust privacy protection and high data utility.

4.3 System Performance

The proposed privacy-preserving dataset sharing scheme introduces minimal overhead and demonstrates lower computational complexity compared to existing methods. Utilizing an efficient perturbation technique based on the XOR mechanism, the scheme significantly reduces time complexity by calibrating noise through column-wise correlation of SNPs, thus expediting the perturbation process without compromising privacy guarantees. This enhancement allows the scheme to handle large genomic datasets practically within a reasonable timeframe.

The scalability of the proposed dataset-sharing solution, when integrated with Lynx.MD, is highly promising. Designed to unlock the power of comprehensive healthcare data, Lynx.MD's robust infrastructure enables efficient management and distribution of large-scale genomic datasets. The integration leverages advanced data

processing and security capabilities, ensuring that the solution can handle increasing volumes of data without performance degradation. The efficient perturbation and utility restoration mechanisms of the proposed scheme maintain low computational complexity, enhancing scalability as the dataset size grows. Consequently, the combined strengths of the platform and the scalable dataset-sharing solution facilitate effective and secure sharing of vast genomic data, driving significant advancements in healthcare research.

5 Conclusion

In this paper, we have demonstrated the feasibility and practicality of a privacy-preserving framework for collaborative genomic research, addressing the critical need for secure data sharing in the era of personalized medicine. By leveraging advanced privacy-preserving algorithms, our solution ensures robust protection of sensitive genomic data while maintaining its utility for research purposes. The collaboration with Lynx.MD has been instrumental in validating our approach, showcasing how industry partnerships can enhance the deployment of privacy-preserving data platforms. Our method effectively balances the trade-offs between data sharing and privacy, providing stakeholders with transparent tools to assess privacy risks and make informed decisions. Our experimental results confirm that the proposed framework outperforms existing methods in both privacy protection and data utility, highlighting its potential for broader application in genomic research and other fields requiring sensitive data handling. The integration of real-time monitoring and visualization tools further enhances the user experience, promoting more effective and secure collaboration. Future work will focus on refining the privacy-preserving techniques and exploring additional applications in other domains. By continuing to address the unique challenges posed by genomic data, we aim to foster global collaboration and drive significant advancements in personalized medicine and public health.

6 Acknowledgment

We would like to extend our sincere gratitude to our Paper Shepherd, Simson L. Garfinkel, for his invaluable guidance and support. We are also deeply appreciative of the anonymous HealthSec peer reviewers, whose detailed feedback and insights significantly enhanced the quality of this paper. We greatly appreciate the world-class peer review provided by the many established leaders in the field.

We would also like to acknowledge Joel Schwartz for his continued and generous support for this project. His contributions have been instrumental to our progress.

References

[1] [n. d.]. opensnp. https://opensnp.org/.
[2] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 308–318.
[3] Hana Lango Allen, Karol Estrada, Guillaume Lettre, Sonja I Berndt, Michael N Weedon, Fernando Rivadeneira, Cristen J Willer, Anne U Jackson, Sailaja Vedantam, Soumya Raychaudhuri, et al. 2010. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 467, 7317 (2010), 832–838.
[4] Erman Ayday, Emiliano De Cristofaro, J Hubaux, and Gene Tsudik. 2015. The chills and thrills of whole genome sequencing. *IEEE Computer Magazine* (2015).
[5] E. Ayday and M. Humbert. 2017. Inference Attacks against Kin Genomic Privacy. *IEEE Security Privacy* 15, 5 (2017), 29–37. https://doi.org/10.1109/MSP.2017.3681052
[6] E. Ayday, J. L. Raisaro, and J. P. Hubaux. 2013. Personal Use of the Genomic Data: Privacy vs. Storage Cost. In *Proceedings of IEEE Global Communications Conference, Exhibition and Industry Forum (Globecom)*.
[7] E. Ayday, J. L. Raisaro, and J. P. Hubaux. 2013. Privacy-Enhancing Technologies for Medical Tests Using Genomic Data. *(short paper) in Proceedings of 20th Annual Network and Distributed System Security Symposium (NDSS)* (2013).
[8] Erman Ayday, Jean Louis Raisaro, Jean-Pierre Hubaux, and Jacques Rougemont. 2013. Protecting and evaluating genomic privacy in medical tests and personalized medicine. In *Proceedings of the 12th ACM Workshop on Privacy in the Electronic Society*. ACM, 95–106.
[9] E. Ayday, J. L. Raisaro, P. J. McLaren, J. Fellay, and J. P. Hubaux. 2013. Privacy-Preserving Computation of Disease Risk by Using Genomic, Clinical, and Environmental Data. In *Proceedings of USENIX Security Workshop on Health Information Technologies (HealthTech)*.
[10] Pierre Baldi, Roberta Baronio, Emiliano De Cristofaro, Paolo Gasti, and Gene Tsudik. 2011. Countering GATTACA: efficient and secure testing of fully-sequenced human genomes. In *Proceedings of the 18th ACM Conference on Computer and Communications Security*. 691–702.
[11] Marina Blanton, Mikhail J Atallah, Keith B Frikken, and Qutaibah Malluhi. 2012. Secure and efficient outsourcing of sequence comparisons. In *Proceedings of European Symposium on Research in Computer Security*. 505–522.
[12] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. 2017. Practical secure aggregation for privacy-preserving machine learning. In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 1175–1191.
[13] Anthony J. Brookes. 1999. The essence of SNPs. *Gene* 234, 2 (1999), 177–186.
[14] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. 2022. Membership Inference Attacks From First Principles. In *IEEE Symposium on Security and Privacy (SP)*. 1897–1914.
[15] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*. 267–284.
[16] Christopher S Carlson, Mark A Eberle, Mark J Rieder, Qinmei Yi, Leonid Kruglyak, and Deborah A Nickerson. 2004. Mapping complex disease loci in whole-genome association studies. *Nature* 429, 6987 (2004), 446–452.
[17] Christopher A. Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. 2021. Label-Only Membership Inference Attacks. In *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139. 1964–1974.
[18] Peter Claes, Denise K Liberton, Katleen Daniels, Kerri Matthes Rosana, Ellen E Quillen, Laurel N Pearson, Brian McEvoy, Marc Bauchet, Arslan A Zaidi, Wei Yao, et al. 2014. Modeling 3D facial shape from DNA. *PLoS Genetics* 10, 3 (2014).
[19] David Clayton. 2010. On inferring presence of an individual in a mixture: a Bayesian approach. *Biostatistics* 11, 4 (2010), 661–673.
[20] Ortal Dayan, Lior Wolf, Fang Wang, and Yaniv Harel. 2023. Optimizing AI for Mobile Malware Detection by Self-Built-Dataset GAN Oversampling and LGBM. In *2023 IEEE International Conference on Cyber Security and Resilience (CSR)*. 60–65. https://doi.org/10.1109/CSR57506.2023.10224927
[21] Emiliano De Cristofaro, Sky Faber, and Gene Tsudik. 2013. Secure Genomic Testing with Size- and Position-hiding Private Substring Matching. In *Proceedings of the 12th ACM Workshop on Privacy in the Electronic Society*.
[22] Cynthia Dwork and Aaron Roth. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science* 9, 3–4 (2014), 211–407.
[23] Yaniv Erlich and Arvind Narayanan. 2014. Routes for breaching and protecting genetic privacy. *Nature Reviews Genetics* 15, 6 (2014), 409–421.
[24] Stephen E Fienberg, Aleksandra Slavkovic, and Caroline Uhler. 2011. Privacy preserving GWAS data sharing. In *IEEE 11th International Conference on Data Mining Workshops (ICDMW)*. 628–635.
[25] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*. 1322–1333.
[26] Neil Gandal, Tyler Moore, Michael Riordan, and Noa Barnir. 2023. Empirically evaluating the effect of security precautions on cyber incidents. *Computers & Security* 133 (2023), 103380. https://doi.org/10.1016/j.cose.2023.103380
[27] Jane Gitschier. 2009. Inferential genotyping of Y chromosomes in Latter-Day Saints founders and comparison to Utah samples in the HapMap project. *American Journal of Human Genetics* 84, 2 (2009), 251–258.
[28] Melissa Gymrek, Amy L McGuire, David Golan, Eran Halperin, and Yaniv Erlich. 2013. Identifying personal genomes by surname inference. *Science* 339, 6117 (2013), 321–324.
[29] Yaniv Harel, Irad Ben Gal, and Yuval Elovici. 2017. Cyber Security and the Role of Intelligent Systems in Addressing its Challenges. *ACM Trans. Intell. Syst. Technol.* 8, 4, Article 49 (may 2017), 12 pages. https://doi.org/10.1145/3057729
[30] Erika Check Hayden. 2013. Privacy protections: The genome hacker. *Nature* 497 (2013), 172–174.

[31] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David W Craig. 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genetics* 4, 8 (2008).

[32] Mathias Humbert, Kévin Huguenin, Joachim Hugonot, Erman Ayday, and Jean-Pierre Hubaux. 2015. De-anonymizing Genomic Databases Using Phenotypic Traits. *Proceedings on Privacy Enhancing Technologies*, 99–114.

[33] Hae Kyung Im, Eric R Gamazon, Dan L Nicolae, and Nancy J Cox. 2012. On sharing quantitative trait GWAS results in an era of multiple-omics data and the limits of genomic privacy. *American Journal of Human Genetics* 90, 4 (2012), 591–598.

[34] Somesh Jha, Louis Kruger, and Vitaly Shmatikov. 2008. Towards practical privacy for genomic computation. In *Proceedings of IEEE Symposium on Security and Privacy*. 216–230.

[35] Tingting Ji, Peng Li, Emre Yilmaz, Erman Ayday, Yanfang Ye, and Jie Sun. 2021. Differentially private binary-and matrix-valued data query: An XOR mechanism. *Proceedings of the VLDB Endowment* 14, 5 (2021), 849–862.

[36] Yuzhou Jiang, Tianxi Ji, Pan Li, and Erman Ayday. 2023. Reproducibility-Oriented and Privacy-Preserving Genomic Dataset Sharing. arXiv:2209.06327 [cs.CR]

[37] Yuzhou Jiang, Tianxi Ji, Pan Li, and Erman Ayday. 2024. Privacy-Preserving Sharing of Genomic Datasets for Research Outcome Validation. *arXiv preprint arXiv:2209.06327v5* (Aug 2024). https://arxiv.org/abs/2209.06327v5.

[38] Aaron Johnson and Vitaly Shmatikov. 2013. Privacy-preserving data exploration in genome-wide association studies. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1079–1087.

[39] Gulce Kale, Erman Ayday, and Öznur Tastan. 2017. A utility maximizing and privacy preserving approach for protecting kinship in genomic databases. *Bioinformatics* 34, 2 (2017).

[40] M. Kantarcioglu, Wei Jiang, Ying Liu, and B. Malin. 2008. A Cryptographic Approach to Securely Share and Query Genomic Sequences. *IEEE Transactions on Information Technology in Biomedicine* 12, 5 (2008), 606–617.

[41] Manfred Kayser and Peter de Knijff. 2011. Improving human forensics through advances in genetics, genomics and molecular biology. *Nature Reviews Genetics* 12, 3 (2011), 179–192.

[42] Z. Lin, A. B. Owen, and R. B. Altman. 2004. Genomic research and human subject privacy. *Science* 305, 5681 (Jul 2004), 183.

[43] Christoph Lippert, Riccardo Sabatini, M. Cyrus Maher, Eun Yong Kang, Seunghak Lee, Okan Arikan, Alena Harley, Axel Bernal, Peter Garst, Victor Lavrenko, Ken Yocum, Theodore Wong, Mingfu Zhu, Wen-Yun Yang, Chris Chang, Tim Lu, Charlie W. H. Lee, Barry Hicks, Smriti Ramakrishnan, Haibao Tang, Chao Xie, Jason Piper, Suzanne Brewerton, Yaron Turpaz, Amalio Telenti, Rhonda K. Roby, Franz J. Och, and J. Craig Venter. 2017. Identification of individuals by trait prediction using whole-genome sequencing data. *Proceedings of the National Academy of Sciences* (2017). https://doi.org/10.1073/pnas.1711125114

[44] Fan Liu, Fedde van der Lijn, Claudia Schurmann, Gu Zhu, M Mallar Chakravarty, Pirro G Hysi, Andreas Wollstein, Oscar Lao, Marleen de Bruijne, M Arfan Ikram, et al. 2012. A genome-wide association study identifies five loci influencing facial morphology in Europeans. *PLoS Genetics* 8, 9 (2012).

[45] Bradley A. Malin and Latanya Sweeney. 2004. How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. *Journal of Biomedical Informatics* 37, 3 (2004), 179–192.

[46] Alisa K Manning, Marie-France Hivert, Robert A Scott, Jonna L Grimsby, Nabila Bouatia-Naji, Han Chen, Denis Rybin, Ching-Ti Liu, Lawrence F Bielak, Inga Prokopenko, et al. 2012. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nature Genetics* 44, 6 (2012), 659–669.

[47] Riley McDowell. 2016. *Genomic selection with deep neural networks.* Master's thesis. Iowa state university.

[48] Priyanka Nanayakkara, Mary Anne Smart, Rachel Cummings, Gabriel Kaptchuk, and Elissa M Redmiles. 2023. What are the chances? explaining the epsilon parameter in differential privacy. In *32nd USENIX Security Symposium (USENIX Security 23)*. 1613–1630.

[49] Milad Nasr, Reza Shokri, and Amir Houmansadr. 2018. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*. 634–646.

[50] Muhammad Naveed, Shashank Agrawal, Manoj Prabhakaran, XiaoFeng Wang, Erman Ayday, Jean-Pierre Hubaux, and Carl Gunter. 2014. Controlled Functional Encryption. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*.

[51] Muhammad Naveed, Erman Ayday, Ellen W Clayton, Jacques Fellay, Carl A Gunter, Jean-Pierre Hubaux, Bradley A Malin, and XiaoFeng Wang. 2015. Privacy in the genomic era. *ACM Computing Surveys (CSUR)* 48, 1 (2015), 6.

[52] Xue-ling Ou, Jun Gao, Huan Wang, Hong-sheng Wang, Hui-ling Lu, and Hong-yu Sun. 2012. Predicting human age with bloodstains by sjTREC quantification. *PloS ONE* 7, 8 (2012).

[53] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. 2018. Scalable private learning with pate. *arXiv preprint arXiv:1802.08908* (2018).

[54] Jean L Raisaro, Florian Tramer, Ji Zhanglong, Diyue Bu, Yongan Zhao, Knox Carey, David Lloyd, Heidi Sofia, Dixie Baker, Paul Flicek, Suyash S Shringarpure, Carlos D Bustamante, Suang Wang, Xiaoqian Jiang, Lucila Ohno-Machado, Haixu Tang, XiaoFeng Wang, and Jean-Pierre Hubaux. 2016. Addressing Beacon Re-Identification Attacks: Quantification and Mitigation of Privacy Risks. *The Journal of the American Medical Informatics Association* 24, 4 (2016), 799–805.

[55] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. 2000. The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision* 40, 2 (2000), 99–121. https://doi.org/10.1023/A:1026543900054

[56] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*. IEEE, 3–18.

[57] Suyash S Shringarpure and Carlos D Bustamante. 2015. Privacy risks from genomic data-sharing beacons. *The American Journal of Human Genetics* 97, 5 (2015), 631–646.

[58] Latanya Sweeney, Akua Abu, and Julia Winn. 2013. Identifying participants in the personal genome project by name. *arXiv preprint arXiv:1304.7605* (2013).

[59] Amalio Telenti, Erman Ayday, and Jean Pierre Hubaux. 2014. On genomics, kin, and privacy. *F1000Research* (Mar 2014). https://doi.org/10.12688/f1000research.4089

[60] Juan Ramón Troncoso-Pastoriza, Stefan Katzenbeisser, and Mehmet Celik. 2007. Privacy preserving error resilient DNA searching through oblivious automata. *Proceedings of ACM CCS '07* (2007).

[61] Nora von Thenen, Erman Ayday, and A. Ercument Cicek. 2018. Re-identification of individuals in genomic data-sharing beacons via allele inference. *Bioinformatics* 35, 3 (2018).

[62] Susan Walsh, Fan Liu, Kaye N Ballantyne, Mannis van Oven, Oscar Lao, and Manfred Kayser. 2011. IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. *Forensic Science International: Genetics* 5, 3 (2011), 170–180.

[63] Rui Wang, Yong Fuga Li, XiaoFeng Wang, Haixu Tang, and Xiaoyong Zhou. 2009. Learning your identity and disease from research papers: information leaks in genome wide association study. In *Proceedings of the 16th ACM Conference on Computer and Communications Security*. 534–544.

[64] Xiao Shaun Wang, Yan Huang, Yongan Zhao, Haixu Tang, XiaoFeng Wang, and Diyue Bu. 2015. Efficient genome-wide, privacy-preserving similar patient query based on private edit distance. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. 492–503.

[65] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *IEEE 31st computer security foundations symposium (CSF)*. 268–282.

[66] Fei Yu, Stephen E Fienberg, Aleksandra B Slavković, and Caroline Uhler. 2014. Scalable privacy-preserving data sharing methodology for genome-wide association studies. *Journal of Biomedical Informatics* 50 (2014), 133–141.

[67] Xiaoyong Zhou, Bo Peng, Yong Fuga Li, Yangyi Chen, Haixu Tang, and XiaoFeng Wang. 2011. To release or not to release: Evaluating information leaks in aggregate human-genome data, In Computer Security – ESORICS 2011 (Leuven, Belgium). *Proceedings of ESORICS'11*, 607–627.

[68] Ligeng Zhu, Zhijian Liu, and Song Han. 2019. Deep leakage from gradients. In *Advances in neural information processing systems*, Vol. 32.

[69] Dmitry Zubakov, Fan Liu, MC Van Zelm, J Vermeulen, BA Oostra, CM Van Duijn, GJ Driessen, JJM Van Dongen, Manfred Kayser, and AW Langerak. 2010. Estimating human age from T-cell DNA rearrangements. *Current Biology* 20, 22 (2010), R970–R971.