Privacy-Preserving Techniques for Genotype-Phenotype Association Studies

Yoolim Annie Choi

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2025

© 2025

Yoolim Annie Choi

All Rights Reserved

Abstract

Privacy-Preserving Techniques for Genotype-Phenotype Association Studies

Yoolim Annie Choi

Genotype-phenotype association studies play a central role in precision medicine by enabling researchers to identify genetic variants that in uence complex traits and diseases with applications ranging from disease risk prediction to treatment strati cation. Notable approaches include genome-wide association studies (GWAS), quantitative trait loci (QTL) mapping, and increasingly, machine learning models that learn predictive relationships between genotype and phenotype data. The power of these approaches grows substantially in collaborative research settings, where pooling data across cohorts increases the statistical power for variant discovery, and sharing trained machine learning models enables institutions with limited computational resources to perform inference on their own cohorts. However, these collaborative gains come with a fundamental privacy-utility tradeoff: sharing raw genomic and phenotypic data can expose sensitive information about study participants, and sharing machine learning models introduces the risk of information leakage through model parameters. These concerns are ampli ed by strict data governance policies and institutional silos that restrict data movement, underscoring the urgent need for privacy-preserving frameworks to enable secure, collaborative analysis while maintaining statistical rigor and practical scalability.

This dissertation develops end-to-end, cryptography-based solutions for privacy-preserving genotype-phenotype association studies, and addresses the entire analysis pipeline from preprocessing and harmonization of phenotype data to secure multi-site eQTL mapping to inference

on sensitive data using machine learning models. The overarching goal is to design frameworks that are designed to (i) provide strong cryptographic security guarantees, (ii) be computationally scalable to large multi-institution cohorts, and (iii) be statistically robust by preserving accuracy comparable to non-secure baselines.

In Chapter 3, we design a privacy-preserving gene expression preprocessing framework that addresses key bottlenecks in federated transcriptomic studies. Our framework provides two secure normalization options — quantile normalization (QN) and relative log expression (RLE) — to allow exibility depending on data sharing standards, a secure multiparty computation (MPC)-based protocol for inverse normal transformation, and a scalable local principal component analysis (PCA)-based hidden covariate correction strategy. We validate our approach using both simulated multi-institution datasets and real-world gene expression data to show that our methods achieve phenotype correction accuracy comparable to centralized, non-secure pipelines while maintaining privacy of individual-level data. These results demonstrate that federated preprocessing with local computation is feasible and effective for collaborative studies.

Building on this foundation, Chapter 4 introduces privateQTL, a secure and scalable framework for multi-center eQTL mapping. privateQTL implements practical genotype and phenotype correction strategies, including genotype population strati cation via projection on public reference panels and the privacy-preserving gene expression preprocessing discussed in Chapter 3. We further propose a one-shot matrix multiplication approach that enables ef cient nominal association testing and permutation-based false discovery control without repeated communication rounds, signi cantly reducing runtime. Our evaluation compares privateQTL against meta-analyses and centralized pipelines across multiple axes — eGene and eVariant discovery rates, robustness to batch effects, statistical power, runtime, and memory footprint — using both simulated federated datasets and real-world multi-site data with known batch heterogeneity. Our results demonstrate that privateQTL achieves superior discovery rates compared to meta-analyses, particularly in heterogeneous data settings, while maintaining strong privacy guarantees under a semi-honest adversary model.

Finally, in Chapter 5, we address the emerging challenge of using secure machine learning inference for sensitive genomic data. We present two HE-based frameworks: (i) a secure inference method for linear models where both inputs and model weights are encrypted, enabling end-to-end con dential inference without compromising predictive performance, and (ii) a method for secure inference on transformer architectures using approximations for non-linear functions. For linear model inference, we introduce an ef cient encoding method that improves computational ef ciency during encrypted dot product computation, and for transformer inference, we develop polynomial approximations for nonlinear functions such as softmax, ReLU, and layer normalization to balance computational feasibility with model accuracy. We validate our linear model inference framework on both continuous and binary phenotype prediction tasks using simulated and real data, achieving performance comparable to plaintext inference. For transformer inference, we discuss the challenges of implementing our approximations in a practical and scalable setting for large scale transformer inference and lay the grounds for future work.

Collectively, this dissertation makes signi cant contributions to the eld of privacy-preserving biomedical informatics. By providing scalable, modular, and cryptographically sound methods for phenotype preprocessing, federated eQTL mapping, and secure machine learning inference, this work enables collaborative genomic research while rigorously protecting sensitive participant data. The frameworks and ndings presented here create a foundation for future developments in privacy-aware collaborative studies, advancing the realization of precision medicine in a manner that respects individual privacy, complies with regulatory requirements, and preserves scienti c reproducibility.

Table of Contents

Acknowledgments ix

Chapter 1: Introduction 1

1.1 Problem Statement 1

1.1.1 Genotype-Phenotype Studies in Biomedicine 1

1.1.2 Privacy Challenges in Large-Scale Genotype–Phenotype Association Studies 2

1.1.3 Privacy-Preserving Approaches for Genotype-Phenotype Association Studies 2

1.2 Objective and Rationale 4

1.3 Research Questions and Hypotheses 5

1.3.1 Aim 1. Methods for Privacy-Preserving Gene Expression Preprocessing in Federated Environments 5

1.3.2 Aim 2. Develop a Scalable Pipeline for Secure and Federated eQTL mapping 6

1.3.3 Aim 3. Develop privacy-preserving genotype-phenotype machine learning inference methods for omics data 8

1.4 Signi cance 9

Chapter 2: Background and Related Works 10

2.1 Overview 10

2.2 Genetic Privacy 11

2.3 Genotype-phenotype association studies 12

2.3.1 Genotype Input for Association Studies 13
2.3.2 Gene Expression Phenotype for Association Studies 14
2.3.3 eQTL mapping 16
2.4 Privacy Concerns in Collaborative Biomedical Research 18
2.4.1 Privacy Concerns in Data Sharing for Federated Studies 18
2.4.2 Privacy Concerns in Model Sharing for Collaboration 18
2.5 Approaches to Mitigating Data Privacy Risks 19
2.5.1 Homomorphic Encryption 20
2.5.2 Secure Multiparty Computation 22
2.6 Prior Work in Privacy-Preserving Genotype-Phenotype Association Studies 23
2.6.1 Meta-analysis for Federated Studies 23
2.6.2 Cryptographic Approaches 24
2.7 Discussion of Gaps 26

Chapter 3: Privacy-Preserving Gene Expression Preprocessing 28
3.1 Overview 28
3.2 Experimental Setup and MPC with Secret Sharing 28
3.3 Implementation Choices for Phenotype Preprocessing 29
3.3.1 Validation of implementation choices 30
3.4 Privacy-preserving phenotype preprocessing methods 31
3.4.1 Privacy-Preserving Quantile Normalization 31
3.4.2 Privacy-Preserving Relative Log Normalization 32
3.4.3 Privacy-Preserving Inverse Normal Transform 32

3.4.4 Privacy-Preserving Phenotype Covariate Correction 36

3.5 Simulated Study 36

3.6 Real-world data 40

3.7 Discussion 42

Chapter 4: Scalable Pipeline for Secure and Federated eQTL mapping 45

4.1 Overview 45

4.2 Privacy-Preserving Genotype Population Strati cation Correction 47

4.3 privateQTL: Secure and Federated cis-eQTL mapping 49

4.3.1 Center and normalizing for cis-eQTL mapping 49

4.3.2 One-shot matrix multiplication for nominal and permutation runs 50

4.4 Dataset description 52

4.5 Benchmarking privateQTL with other eQTL mapping techniques 53

4.6 privateQTL Delivers Accurate Results with Federated GTEx Data 54

4.6.1 eGenes Discovery Analysis 54

4.6.2 Comparison with LIMIX 61

4.6.3 eGene-eVariant Analysis 61

4.6.4 Effect Size Analysis 63

4.6.5 MPC Modulo Operation Accuracy Assessment 65

4.7 Demonstration of privateQTL Capabilities in a Real-World Setting with Batch Effects Highlights its Practical Adaptability 67

4.7.1 eGenes Discovery Analysis 67

4.7.2 Meta-Analysis vs. PrivateQTL in eQTL Mapping: A Comparative Analysis 70

4.8 PrivateQTL is Practical and Scalable 74

4.9 Security Analysis of privateQTL 76

4.9.1 Security Assumption 76

4.9.2 Software Security 77

4.9.3 Security of privateQTL-I 77

4.9.4 Security of privateQTL-II 80

4.10 Discussion 81

Chapter 5: Privacy-Preserving Genotype-Phenotype Machine Learning Inference Methods for Omics Data 83

5.1 Overview 83

5.2 Secure Inference with Encrypted Model Weights 83

5.2.1 Dataset 84

5.2.2 Plaintext Training 85

5.2.3 Secure Inference Model Implementation 91

5.2.4 Secure Inference Results 94

5.3 Secure Inference for Transformers 100

5.3.1 Dataset 101

5.3.2 Embedding layer 102

5.3.3 Positional Encoding 102

5.3.4 Transformer Block 102

5.3.5 Considerations for HE implementation 103

5.3.6 Consolidating Approximations in Secure Inference 110

Chapter 6: Conclusions and Future Work 111

6.1 Privacy-preserving gene expression preprocessing 111

6.1.1 Contributions 111

6.1.2 Limitations 112

6.1.3 Recommendations for Future Work 112

6.2 Scalable pipeline for secure and federated eQTL mapping 113

6.2.1 Contributions 113

6.2.2 Limitations 114

6.2.3 Recommendations for Future Work 115

6.3 Privacy-preserving genotype-phenotype machine learning inference methods for omics data 115

6.3.1 Contributions 115

6.3.2 Limitations 116

6.3.3 Recommendations for Future Work 116

6.4 Contribution to Biomedical Informatics 117

References 119

Appendix A: Estimation of runtime for multi-layer NNs with homomorphic encryption . . 130

List of Figures

3.1 Privacy-preserving phenotype preprocessing methods. 35

3.2 Effect of gene expression pre-processing on QTL mapping performance 40

3.3 Batch effects before and after each covariate correction method 42

4.1 Schematic overview of privateQTL 46

4.2 Implementation details of privateQTL mapping 52

4.3 privateQTL results for federated GTEx data 57

4.4 Distance between top eVariant and gene TSS as a function of corrected p-value for federated GTEx data 57

4.5 privateQTL-I eGenes false positives and false negatives compared to GTEx pipeline 59

4.6 privateQTL-II eGenes false positives and false negatives compared to GTEx pipeline 60

4.7 privateQTL result compared to LIMIX 62

4.8 p-value landscape for signi cant eQTLs of egenes HSPA12B, C1orf145, and SSSCA1 64

4.9 p-value landscape for signi cant eQTLs of egenes SRD5A3, MAGEF1, and RP11-6N17.1 64

4.10 Accuracy comparison with plaintext versions of privateQTL-I and privateQTL-II . 66

4.11 privateQTL results for federated real-world data with batch effects 69

4.12 Distance between top eVariant and gene TSS as a function of corrected p-value for federated real-world data with batch effects 69

4.13 eGenes found by GTEx pipeline and privateQTL but not Meta-analysis 71

4.14 Scalability analysis of privateQTL runtime with increased sample size 76

5.1 Illustration of our secure inference scenario with encrypted model weights 84

5.2 Comparison of feature selection methods for each linear model 87

5.3 Model performance with different levels of feature selection 88

5.4 Neural Network performance with differing layers and activation function 89

5.5 The process of encoding, encryption, and evaluation of test data and trained linear model 97

5.6 Accuracy in AUC and AUPRC for simulated and real dataset for linear model inference 98

5.7 The comparison between true and securely predicted values for secure continuous phenotypes on the test dataset 98

5.8 Time estimation of computing the rst fully-connected layer of a multi-layer NN with HE. 99

5.9 Illustration of our secure inference scenario for transformer models 100

List of Tables

3.1 Evaluation of different phenotype preprocessing choices 31

3.2 Simulation strategies for phenotype pre-processing comparison 38

4.1 Evaluation of genotype preprocessing choices on eQTL mapping eGene discovery . 48

4.2 Overlap of eGenes compared to GTEx pipeline from combination of sites 72

4.3 Comprehensive summary of privateQTL compared to Meta-analysis 75

5.1 Tuned parameters for Multilayer Neural Networks 90

5.2 Performance of the models including the roundtrip time 95

5.3 Estimation of time cost for the rst fully-connected layer using CKKS scheme . . . 96

Acknowledgements

Very appropriately, this dissertation—on the importance of collaboration—is itself the result of the guidance, support, and contributions of many people. Firstly, I would like to thank my advisor Dr. Gamze Gürsoy. Thank you for supporting my exploration of research questions and for believing in me to see my projects through to completion. You have been the most devoted supporter for my growth and success, and without your guidance, I would not have made it this far.

I'd like to thank my committee members–Dr. Yufeng Shen, Dr. Itsik Pe'er, and Dr. David Knowles and Dr. Peihan Miao. Thank you for your mentorship and guidance throughout my PhD journey. Thank you Dr. Peihan Miao, Dr. Tuuli Lappalainen, Dr. Seungwan Hong, Yebin Kim, Dr. Kirill Nikitin, and Jacob Blindenbach–it was a privilege to have worked with you.

Thank you to my fellow G2 lab members, especially Carlos, Rashid, Xiaoting, Yebin, Hannah, Kae, Ahmed, Conor, and Anoushka, for lling my non-research times with a healthy combination of peace and excitement, and for being my support system through uncertain times.

To my parents–I now understand the journey you have taken, accomplishing your own academic career whilst having a family. To my brother–thank you for understanding my thoughts and needs without words. I deeply cherish our silent companionship.

To my Sadie–your soft droopy ears, puppy dog eyes, and morning whines were essential to power through during tough times. To my husband, Tom–you inspire me to be con dent and fearless, to believe in my capabilities, and to push further. Thank you for being my rock. I love you endlessly.

Chapter 1: Introduction

1.1 Problem Statement

1.1.1 Genotype-Phenotype Studies in Biomedicine

Characterizing genotype–phenotype relationships is fundamental to the advancement of personalized medicine and the development of targeted therapeutic interventions [1, 2]. These relationships can be investigated through several approaches. Genome-wide association studies (GWAS) identify statistical associations between genetic variants and disease phenotypes, whereas quantitative trait loci (QTL) mapping characterizes the relationship between genetic variation and quantitative molecular traits such as gene expression (eQTL) or alternative splicing (sQTL) [3, 4, 5]. In addition, machine learning–based approaches have been employed to capture various genotype–phenotype associations using exible modeling frameworks [6, 7, 8, 9]. Recent advancements, particularly with transformer-based models like DNABERT and the Nucleotide Transformer, have further enhanced these capabilities by extracting deep semantic information from genomic sequences, moving beyond simple regulatory landscape interpretation [6, 10].

Both GWAS and QTL mapping require large sample sizes to achieve suf cient statistical power, particularly when analyzing rare tissue types or uncommon genetic variants [11, 12]. As a result, these studies bene t greatly from large-scale collaborative efforts and multi-center data sharing initiatives. The recent surge in sequencing technologies presents an opportune moment to conduct large-scale cross-institutional studies. In parallel, machine learning models have emerged as powerful tools for genotype–phenotype prediction, but training these models often requires substantial data and computational resources that may not be available at all research centers. Consequently, many institutions rely on pretrained models developed by larger centers and primarily perform in-

ference on their own data. Collaborative efforts across research sites—whether through pooled data for GWAS and QTL mapping or shared model resources for machine learning—hold signi cant potential to advance genotype–phenotype association studies and accelerate biological discovery.

1.1.2 Privacy Challenges in Large-Scale Genotype–Phenotype Association Studies

The pursuit of collaborative research often con icts with the imperative to protect patient privacy. Sensitive information can be inadvertently exposed through multiple avenues, including raw genetic data, aggregated summary statistics, and functional genomics measurements such as gene expression, which can serve as a proxy for an individual's genomic pro le [13, 14, 15, 16, 17]. Growing reports of information leakage, even from single-cell gene expression data [18], underscore the need for comprehensive privacy protection across all data types.

These privacy risks extend beyond data sharing to the exchange of trained machine learning models. When models are shared with other centers for inference, sensitive information may still be inferred from model weights or predictions. Model providers may be reluctant to share their models due to the risk of information leakage [19, 20, 21], while clients seeking to use the model for inference may hesitate to disclose sensitive input data to the model owner or a third-party server. This scenario highlights the pressing need for privacy-preserving approaches to machine learning inference that protect both model parameters and client inputs.

In addition, various institutional regulations on data sharing further complicate research cooperation by requiring researchers to navigate regulatory frameworks while ensuring data protection throughout the research process [22, 23].

1.1.3 Privacy-Preserving Approaches for Genotype-Phenotype Association Studies

Cryptographic techniques such as homomorphic encryption (HE) or secure multiparty computation (MPC) can be potential solutions to these privacy concerns, as they provide theoretical guarantees on data security [24, 25, 26]. HE ensures data security by encoding information into a

polynomial space, allowing computations to be performed directly on the encrypted data without revealing the underlying plaintext–a process referred as homomorphic evaluation. As HE relies on polynomial approximations of non-linear operations, their application becomes dif cult for complex biomedical analyses, and the extensive time required for these operations further complicates their use. MPC, on the other hand, relies on joint computations among servers that each hold partial information about the private input data. While MPC avoids the noise management and computational complexity challenges associated with HE, it introduces inherent communication overhead between the participating servers. Effectively integrating these cryptographic techniques into bioinformatics work ows requires dual innovation: adapting cryptography to the complexities of biomedical data while also tailoring bioinformatics methodologies to the constraints of secure computation.

Prior studies have shown promise in using these cryptographic techniques on association studies, particularly for GWAS[27, 28]. GWAS nds associations between genetic variants and a disease phenotype of interest, where the genotype is represented as a matrix, and the phenotype as a vector. However, as multi-omics data becomes increasingly prevalent, a research gap remains in privacy-preserving solutions for other types of association studies, such as QTL mapping and integrative multi-omics analyses. Employing privacy preserving techniques for QTL mapping is challenging, as it involves analyzing correlations between tens of thousands of phenotypic traits and millions of genetic variants. In addition to the scalability concerns posed by the sheer volume of input data, phenotype data such as gene expression also introduces batch effects and noise due to variations in data collection across multiple sites.

Previous studies have investigated the use of HE for various machine learning applications [29, 30, 31]. However, research gaps remain, such as in evolving privacy protection scenarios that safeguard not only input data but also model weights. With the widespread adoption of novel architectures such as transformers, there is a growing need for secure inference methods capable of supporting these more complex models.

Addressing the challenges of privacy-preserving genotype-phenotype association studies re-

quires solutions that balance scalability, data harmonization, and model privacy. As multi-center collaborations become essential for enhancing statistical power, ensuring proper data harmonization is critical to mitigate biases arising from heterogeneous data sources. Additionally, the large-scale nature of QTL mapping demands computationally ef cient methods capable of handling vast genetic and phenotypic datasets without compromising accuracy. Finally, with the rapid evolution of machine learning in biomedical research, secure inference techniques must adapt to protect model weights and input data while keeping pace with emerging architectures.

1.2 Objective and Rationale

An effective approach to utilize cryptographic techniques for genotype-phenotype association studies must incorporate privacy-preserving measures throughout the entire research pipeline—from data preprocessing and harmonization to statistical analysis and post-training inference. Additionally, such an approach must be scalable and generalizable in diverse settings while maintaining accuracy comparable to non-secure methods to ultimately enable collaborative studies to achieve their full potential without compromising patient privacy.

Building on these challenges, this thesis seeks to address barriers to the practical implementation of privacy-preserving methodologies, supporting their adoption in real-world genotype-phenotype association studies. To this end, we present methods designed to enhance privacy protection for various research scenarios across the genotype-phenotype association work ow. Speci cally, we focus on three key areas: developing secure preprocessing methods for genotype and phenotype data, establishing a robust pipeline for privacy-preserving multi-center QTL mapping, and exploring secure techniques for the evaluation of machine learning models that use private genetic data. We focus on eQTL mapping as a speci c genotype-phenotype association study, where scalability is crucial to accommodate large-scale, federated input data. In addition, this thesis addresses privacy-preserving machine learning inference, with an emphasis on safeguarding sensitive information across diverse inference scenarios and adapting to contemporary machine learning architectures.

1.3 Research Questions and Hypotheses

1.3.1 Aim 1. Methods for Privacy-Preserving Gene Expression Preprocessing in Federated Environments

i. How can gene expression normalization methods be designed to preserve analytical utility while protecting individual privacy?

ii. What approaches can be used to develop privacy-preserving methods for inverse normal transformation and hidden covariate correction while minimizing information leakage?

Experimental Design for Aim 1. Aim 1 focuses on developing methods for gene expression preprocessing in multi-center, federated environments, where data are horizontally partitioned across institutions. This distribution poses unique challenges: differences between centers introduce batch effects, and hidden covariates can confound downstream analyses. Existing phenotype preprocessing methods typically require access to the complete dataset, which raises signi cant data privacy concerns. To address these challenges, we design a framework that enables privacy-preserving normalization, transformation, and hidden covariate correction of gene expression data while maintaining analytical utility. We prioritize achieving accuracy on par with aggregated preprocessing methods while maintaining computational ef ciency and practical applicability.

We develop privacy-preserving methods for two widely used gene expression normalization techniques—quantile normalization (QN) and relative log expression (RLE). The design principle for our methods is to maximize local computation, ensuring that only non-sensitive aggregate information is shared with computing parties while maximizing computation ef ciency. Following normalization, we transform the gene expression values to follow a standard normal distribution by introducing a novel multiparty computation (MPC)-based protocol for inverse normal transformation across samples. Our protocol ensures that no private information about individual expression values is revealed during computation while maintaining statistical accuracy. We optimize our method to enable ef cient parallelization of MPC operations, improving overall computational

performance. Finally, we apply a local principal component analysis (PCA)-based correction on the normalized and transformed expression values, which effectively removes residual site-speci c hidden covariates while minimizing information sharing between sites. Together, these methods constitute a practical and privacy-preserving preprocessing framework that enables collaborative transcriptomic analyses without compromising data con dentiality.

We validate our privacy-preserving preprocessing framework using both simulated and real-world multi-center gene expression datasets. Simulated data allows quantitative assessment of how well the method corrects confounding factors, since the true covariates are known and can be used as a ground truth for evaluation.We also use real-world federated datasets where data was collected and quanti ed in three different centers. This data is used to visualize the impact of preprocessing on batch effects across centers, assessing how well our privacy-preserving methods correct batch effects compared to the standard aggregated pipeline and raw federated data.

1.3.2 Aim 2. Develop a Scalable Pipeline for Secure and Federated eQTL mapping

i. What strategies enable the design of a secure eQTL mapping framework that supports cross-institutional collaboration while safeguarding individual privacy?

ii. To what extent does our tool achieve greater accuracy compared to current standards for federated eQTL mapping?

Experimental Design for Aim 2. Aim 2 builds on the phenotype preprocessing methods developed in Aim 1 and focuses on constructing privacy-preserving pipelines for genotype–phenotype association studies, speci cally expression quantitative trait loci (eQTL) mapping. We introduce privateQTL, an MPC-based framework for secure and scalable cis-eQTL mapping in federated settings [32]. The goal of this aim is twofold: (1) to design computationally ef cient methods for large-scale association testing that preserve privacy, and (2) to rigorously evaluate privateQTL against state-of-the-art approaches.

Before performing eQTL mapping, both genotype and phenotype data must undergo prepro-

cessing to ensure that downstream association analyses are accurate and reliable. This preprocessing typically includes steps such as quality control, normalization, and correction for confounding factors like population structure or hidden batch effects. While these procedures are essential for robust association testing, they are not the focus of this thesis.

Following preprocessing and covariate correction, the genotype and phenotype matrices are subjected to eQTL mapping. Current approaches involve repeated rounds of matrix multiplication for the nominal run (signal detection) and multiple permutation runs (null distribution generation), which can be computationally expensive in a secure setting. (Please refer to Chapter 2 section 2.3.3 for details.) To address this limitation, we develop a one-shot matrix multiplication strategy that computes all necessary association statistics for a gene in a single pass, signi cantly reducing computational overhead. Furthermore, we parallelize the privateQTL pipeline to ensure scalability to large-scale cohorts, making the method practical for real-world deployment.

We evaluate privateQTL under two privacy-preserving scenarios to benchmark its performance. First, we use an aggregated setting, where all data are centralized and processed jointly, providing a gold-standard reference for performance. Second, we use meta-analysis, the current standard for federated studies, where each site independently performs eQTL mapping and shares only summary statistics for downstream aggregation. Validation is performed on two datasets: (i) a simulated federated dataset, in which data from one center is arti cially partitioned into three, and (ii) a real-world federated dataset, where expression and genotype data are collected and processed independently across three centers, introducing realistic heterogeneity. We assess privateQTL's performance across ve dimensions: (1) eQTL discovery accuracy relative to the aggregated benchmark, (2) biological interpretability of top eQTLs, (3) robustness to inter-site heterogeneity, (4) scalability to large sample sizes, and (5) runtime and memory ef ciency. Together, these experiments provide a comprehensive evaluation of the accuracy, utility, and practicality of privateQTL for secure eQTL mapping in federated settings.

1.3.3 Aim 3. Develop privacy-preserving genotype-phenotype machine learning inference methods for omics data

i. How can inference methods for linear models be developed to ensure that both input data and model weights remain private?

ii. What approaches allow the development of privacy-preserving methods for evaluating transformer models while ensuring reliable model performance?

Experimental Design for Aim 3. Aim 3 focuses on developing privacy-preserving genotype–phenotype machine learning inference methods for omics data. Our overarching goal is to develop homomorphic encryption–based inference frameworks that (1) protect both the client's input data and the model's parameters from disclosure, and (2) enable secure evaluation of complex models, including architectures such as transformers. We aim to achieve accuracy comparable to plaintext inference while ensuring computational practicality.

We begin by developing a secure inference framework for linear regression, where both model weights and input data are kept private. As linear regression inference consists of computing a dot product between input feature vectors and weight vectors, we use an ef cient input-packing strategy that minimizes cryptographic operations and maximizes computational throughput. Our method is evaluated on both prediction accuracy and computational ef ciency using two datasets representative of real-world use cases.

We next extend the framework to transformer models, which require repeated matrix multiplications and the application of non-linear operations such as softmax, layer normalization, and ReLU. Since homomorphic encryption supports only addition and multiplication, we present ways that can approximate these non-linear operations with polynomials. The objective is to design approximations that preserve predictive accuracy without imposing excessive computational overhead in the context of complex transformer models.

1.4 Signi cance

Integrating vast amounts of patient data is essential for gaining genetic insights and ultimately improving personalized medicine. The advancement of genotype-phenotype association studies, however, must be coupled with privacy-preserving technologies to address the growing concern of information leakage. Yet there is a constant trade-off between utility and privacy, as sensitive information oftentimes contains the most research potential. This thesis addresses a critical need for privacy-preserving methodologies in genotype-phenotype association studies, without compromising research integrity and practical feasibility. Speci cally, this thesis introduces innovative solutions utilizing cryptographic techniques–such as HE and MPC–for association studies, addressing gaps in prior work where scalability limitations restricted the practical application of privacy-preserving tools. As such, this work not only enhances privacy safeguards but also ensures that collaborative studies can scale effectively to large, multi-center datasets. In addition, it examines diverse real-world scenarios in genotype-phenotype association studies, including federated studies aimed at enhancing statistical power and sharing of machine learning models for post-training inference. These scenarios represent all stages of genotype-phenotype association studies, and addresses the researcher's diverse needs for privacy protection. Overall, these ndings have the potential to transform privacy-preserving biomedical research, enabling broader adoption of secure methods in the eld and contributing to the advancement of personalized medicine.

Chapter 2: Background and Related Works

2.1 Overview

This chapter provides the background and related work that form the foundation of this dissertation. It begins with introducing genetic privacy, its importance, and reports of leakage in various levels of summarized information, as described in section 2.2. The chapter then presents an overview of genotype–phenotype association studies (section 2.3), describing widely used methods for uncovering genetic associations, including genome-wide association studies (GWAS), expression quantitative trait locus (eQTL) mapping, and emerging machine learning approaches. This section also details the preprocessing steps required for genotype and phenotype inputs, and provides a step-by-step description of the eQTL mapping process as it pertains to this work (subsection 2.3.3).

Next, the chapter examines privacy concerns in collaborative studies, focusing on institutional collaborations where data sharing and model sharing are increasingly common (section 2.4). These settings introduce distinct privacy challenges, particularly in federated studies and cross-institutional inference. To address such challenges, the chapter surveys privacy-enhancing technologies (PETs), with detailed explanations of homomorphic encryption (HE) and secure multiparty computation (MPC), highlighting their differing privacy guarantees, assumptions, and limitations (section 2.5).

The chapter then reviews prior work in privacy-preserving association studies, including current standards such as meta-analysis approaches and previous attempts to incorporate cryptographic tools for secure genotype–phenotype association (section 2.6). Finally, it identi es gaps in existing literature, outlining the limitations of current methods in maintaining accuracy, scalability, and usability in privacy-preserving genetic analysis. Together, these sections establish the

technical and conceptual foundation upon which this dissertation builds.

2.2 Genetic Privacy

This dissertation addresses the protection of individual genetic information. Unlike other forms of personal data, genetic information is inherently immutable and uniquely identifying. A breach of such data poses direct and irreversible risks to privacy, as an individual's genetic blueprint cannot be changed or replaced in the way a compromised social security number or credit card can. Protection of genetic privacy extends beyond the underlying genotype itself–attributes such as study participation, demographic characteristics, and socioeconomic status can also be inferred from genetic data and therefore warrant equal protection. Such breach in privacy can occur despite protective measures, such as releasing partial or anonymized genetic information for research purposes. One well-known example is the identi cation of James Watson's APOE—gene associated with Alzheimer's disease—risk status, despite the variant having been masked prior to the public release of his genomic data for research purposes[33]. Anonymized genomes, in which direct identi ers have been removed, can nonetheless be cross-referenced against genealogy databases to infer additional identifying details, such as an individual's surname or familial relationships[34, 35].

Additionally, summarized genetic information—such as allele frequencies, association study statistics, or functional genomics measurements that re ect the genome's biological activity—can also pose signi cant risks of privacy leakage. In a seminal 2008 study, Homer et al. showed that allele frequency information can inadvertently reveal an individual's participation in a study cohort, highlighting the vulnerabilities in genomic privacy [36]. This issue is particularly concerning for underrepresented groups, who face a heightened risk of identi cation and resulting potential stigmatization. While functional genomics data that represent a genome's biological activity, such as gene expression, have traditionally been considered not private, recent studies have shown gene expression data can be used to link individuals' genotypes using publicly available datasets that

explain the correlation between genotype and phenotype [37, 38]. These linking attacks exploit publicly available eQTL data to back trace genotype information from released phenotype data. The partial genotypes created from this correlation are suf cient to identify an individual in a sensitive dataset. Furthermore, a recent research reported by our group has shown that these linking attacks extend to single cell RNA-sequencing data using cell-type speci c pseudo-bulking [18].

Privacy issues persist in machine learning as well, with reports of information leakage from machine learning models [39]. A foundational study led by Shokri et al. has shown that access to model weights or query capability of a machine learning model can allow adversaries to infer membership of a sample in the training set by leveraging con dence in model inference [40]. This attack exploits the assumption that over- tted models exhibit high con dence when predicting previously seen data, while displaying lower con dence for novel inputs. As model architectures grow increasingly complex and pre-trained models are widely shared for research, integrating privacy-aware approaches has become a critical consideration. Failure to address privacy issues in genetic data can cause dignitary and monetary harm, such as discrimination in employment and insurance based on racial or medical conditions.

2.3 Genotype-phenotype association studies

Characterizing the relationships between genotypes and phenotypes is essential for developing personalized medicine and therapies [1, 2]. Genome-wide association studies (GWAS) are widely used to nd variants that are correlated with disease phenotypes [41]. GWAS has led to thousands of new discoveries, signi cantly advancing biomedical research and the eld of precision medicine [42]. As large-scale omics projects—which encompass comprehensive pro ling of molecular layers such as genomics, transcriptomics, proteomics, and metabolomics—increasingly generate data from tissue samples, we can now interpret the functional roles of variants that in- uence molecular phenotypes, such as gene expression, through expression quantitative trait loci (eQTL) mapping [43]. eQTL studies nd associations between genotype and gene expression and

can aid with identifying the functional role of implicated variants in GWAS [3, 4]. This type of association can be expanded to other molecular phenotypes such as alternative splicing (sQTL) and methylation (mQTL) [5, 44]. Compared to GWAS, where the phenotype space is a vector that describes a disease phenotype, the phenotype space in QTL mapping is a matrix, with rows that range from 20,000 to 30,000 genes. This expansion requires additional steps for association, the details of which are described in subsection 2.3.3.

Machine learning (ML) provides a powerful framework for estimating genotype-phenotype relationships. Linear regression allows the prediction of continuous traits such as height or heritability, while logistic regression enables the classi cation of disease outcomes [45, 46, 47]. Beyond traditional linear methods, machine learning models are increasingly utilized to capture non-linear interactions between genotypes and phenotypes [48]. ML-based genotype-phenotype association studies include the application for GWAS[49, 50], cancer prediction[51, 52], or prediction of polygenic risk scores (PRS) that provides an aggregate estimate of the risk measure across multiple variants[53, 54]. The development of advanced model architectures, such as transformers, has further expanded the application of complex machine learning approaches in genomics [55, 6, 10].

In the following sections, we provide an overview of the datasets used as genotype and phenotype inputs for association studies, along with the preprocessing steps required to ensure robustness in the subsequent association analyses. Finally, we outline the eQTL mapping process as it relates to the speci c objectives of this dissertation.

2.3.1 Genotype Input for Association Studies

A common type of genomic variation that is examined for genotype-phenotype association studies is single nucleotide polymorphisms (SNP), which refer to variation at a single position in the DNA sequence among individuals. For a typical genotype matrix used in GWAS or QTL mapping studies, the dataset commonly comprises approximately 3 to 5 million SNPs. A critical rst step in enabling secure and reliable analyses is addressing sources of bias that can confound association results. In genotype data, the most prominent confounder is population strati cation,

which refers to systematic genetic differences between subpopulations that are driven by ancestral background rather than true associations with the phenotype of interest [56, 57]. Population strat- i cation is typically detected and corrected using principal component analysis (PCA), where the rst few principal components capture the major axes of genetic variation. When plotted, the rst two principal components often cluster samples according to their population structure. Correction is usually done by residualizing the genotype data with the principal components [58]. Population structure can also be incorporated directly as random effects based on genetic relatedness matrix in linear mixed models (LMM) such as LIMIX [59].

2.3.2 Gene Expression Phenotype for Association Studies

Depending on the research project, multiple types of phenotype inputs may be relevant, including binary disease phenotypes, alternative splicing events, protein abundance, or quantitative gene expression levels. Among these, gene expression is particularly informative, as it captures the transcriptional activity across the genome and provides insight into cellular states and regulatory mechanisms. Historically, gene expression was measured primarily using microarrays; however, advances in next-generation sequencing (NGS) technologies, coupled with decreasing sequencing costs, have made RNA sequencing (RNA-seq) the dominant platform. RNA-seq offers improved dynamic range, higher resolution, and the ability to detect novel transcripts compared to microarrays [60, 61]. A typical RNA-seq dataset used in association studies generally includes between 20,000 and 30,000 genes. The resulting count-based data requires rigorous normalization and transformation to account for sequencing depth, library composition, and technical variability, making preprocessing a critical step in downstream analyses such as eQTL mapping and integrative multi-omics studies.

Gene expression preprocessing is a crucial step in data analysis that removes non-biological variation that may be caused during RNA-sequencing. Failure to remove technical batch effects can result in exaggerated or misleading conclusions in analysis [62]. A standard preprocessing pipeline for gene expression consists of normalization, inverse normal transform, and hidden covariate

correction.

Gene Expression Normalization and Inverse Normal Transform After ltering low read counts, gene expression is normalized to allow gene expression levels to be comparable across samples, removing unwanted biases created during RNA extraction, library preparation, and sequencing. Common normalization methods include quantile normalization (QN), trimmed mean of M values (TMM), and relative log expression (RLE) [63, 64, 65]. QN adjusts across-samples statistical distributions to be the same and requires computing the rank of the genes based on their expression value for each sample and then replacing each rank with their across-sample rank average. TMM and RLE are designed to correct for RNA composition bias, where a small number of highly expressed genes can distort library size estimation. In such cases, global adjustment methods may not be suitable. TMM selects a representative sample from the dataset to serve as a reference for computing size factors. These size factors are derived by comparing the log-fold changes of gene expression in each sample relative to the reference sample, allowing for normalization that accounts for library size and RNA composition biases. RLE calculates the geometric mean of gene expression across samples to serve as a pseudo-reference and determines the library size factor for each sample relative to this reference. RLE is widely used for differential gene expression analysis such as DESeq [66, 67]. While many non-secure pipelines choose TMM as their normalization method [43, 68], a systematic comparison of normalization methods showed that TMM and RLE have equivalent performance in eQTL mapping [69].

The normalized gene expression matrix is transformed to follow a standard normal distribution per gene. This requires ranking of expression values across samples per gene, and converting them into rank percentiles by dividing the ranks by the total number of samples. These rank percentiles are then treated as cumulative distribution function (CDF) values, and converted to their z-score counterpart using inverse normal transform.

Gene Expression Hidden Covariate Correction In a standard gene expression preprocessing pipeline, normalized and transformed gene expression data is then corrected for hidden covariates.

One approach for mitigating these effects is Combat, which applies an empirical Bayes framework to model batch effects as a Gaussian distribution [70]. Combat-seq extends this method by employing a negative binomial distribution instead of a normal distribution, improving batch effect correction for RNA-seq count data [71]. A widely used hidden latent factor estimation method for phenotype data is Probabilistic Estimation of Expression Residuals (PEER), as used in GTEx consortium and Geuvadis study [72, 43, 73]. PEER aims to nd hidden latent factors that are in uenced by batch effects and population structure using a bayesian framework and expectation-maximization (EM) algorithm. PCA is another covariate correction method used for gene expression preprocessing method. Overall, these approaches correct batch effects in a single, aggregated setting, but when data are distributed across multiple centers, privacy concerns and the inability to share raw data make traditional aggregation-based methods impractical.

2.3.3 eQTL mapping

eQTL mapping refers to the process of identifying statistical associations between genetic variants (genotypes) and gene expression levels (phenotypes). There are two main types of eQTL mapping: cis and trans. Cis-eQTL mapping focuses on identifying variants located within a de ned genomic window—typically within 1 megabase (Mb)—of the gene of interest. For each gene, single nucleotide polymorphisms (SNPs) within this window are tested for association with that gene's expression. In contrast, trans-eQTL mapping examines associations between all variants and all genes across the genome, aiming to detect long-range regulatory effects that occur between distant loci.

Before eQTL mapping, both genotype and phenotype data must undergo preprocessing and batch effect correction. Genotype matrix, often consisted of SNPs by individuals, requires population strati cation correction to address population structure as confounder. Gene expression matrix, consisted of genes by individuals, requires removal of non-biological confounders. Genotype correction and gene expression matrix preprocessing can be performed using the methods outlined above.

With the expansion of the phenotype space from a vector in GWAS to a matrix in QTL mapping studies, individual association testing for each SNP variants and gene pair becomes impractical. To address this, Shabalin et al. introduced Matrix eQTL, which enables the computation of associations between SNPs and genes through matrix multiplication [74]. This framework was further improved with faster permutation tests via null distribution modeling and tensor-based computation in fastQTL and tensorQTL, respectively [75, 76]. fastQTL introduced modeling the null distribution as a beta distribution, reducing the number of permutations required by the user to establish the null distribution for false discovery control. As a result, the users are now able to obtain accurate results with around 1000 permutations, instead of 10– 000. Overall, this multiplication-based association allows computing parties in MPC setting to maximize local computation.

Due to the vast number of association tests performed for SNP-gene pairs, eQTL mapping requires permutation modeling to control false discovery rates. For this, the gene expression matrix is shuf ed a thousand times to model a null distribution[75]. This permutation modeling is non-trivial in a secure setting, as it requires repeated MPC-multiplication. Instead, the process can be expedited for cis-eQTL mapping where the gene vector can be shuf ed 1000 times and concatenated to form a phenotype matrix. This phenotype matrix, along with the cis-genotype matrix for the speci c gene vector, can allow one matrix multiplication for both nominal and permutation runs. False discovery rate (FDR) control is applied using a null distribution modeled from permutation runs. Genes that exhibit statistically signi cant associations with genetic variants after this correction are termed `eGenes', and the corresponding variants are referred to as `eVariants'. Each eGene–eVariant pair is characterized by a p-value, which quanti es the signi cance of the association; an effect size, which represents the magnitude and direction of the association; and a standard error, which re ects the uncertainty in the estimated effect size.

2.4 Privacy Concerns in Collaborative Biomedical Research

2.4.1 Privacy Concerns in Data Sharing for Federated Studies

Privacy concerns become pronounced in collaborative genotype-phenotype association studies involving multiple research centers. In one common model of collaboration, institutions share data to jointly conduct association analyses, aiming to increase statistical power through larger sample sizes, enhance population diversity, and improve access to rare or otherwise hard-to-obtain tissue types. GWAS and eQTL studies typically include sample sizes that range from a few hundred to several thousand [11, 12]. Obtaining suf cient statistical power for hard-to-access tissues or rare disease types requires larger sample sizes. To address this issue, researchers often undertake multi-site studies, aggregating data from multiple institutions [77, 78, 79]. Further details on current best practices for federated studies while protecting privacy are outlined in section 2.6. While data sharing mitigates limitations in statistical power, it simultaneously introduces signi cant privacy risks, as sensitive input data must be exchanged between institutions. This is compounded by administrative policies and legislation like the General Data Protection Regulation (GDPR) or the Health Insurance Portability and Accountability Act (HIPAA), which provide legal protection over patient data [22, 23]. In accordance with these regulations, newly generated genetic data are often siloed in data centers with restricted access. As a result, cross-jurisdictional data sharing requires multiple administrative approvals to grant researchers access, and is often limited to a small group which hinders replicability and broader scienti c validation. Collaborative studies that aim to increase statistical power for rare diseases and minority population groups face a unique challenge where their data scarcity necessitates collaboration for research but increases identi cation risks.

2.4.2 Privacy Concerns in Model Sharing for Collaboration

With the advancement of complex machine learning architectures for genotype–phenotype association studies, collaborative research has expanded beyond data sharing to include model shar-

ing. This form of collaboration involves sharing pre-trained machine learning models, rather than conducting joint association studies through direct data sharing.

Training transformer-based machine learning models is an exceptionally resource-intensive process, often beyond the computational capacity of many research institutions. For example, DNABERT takes 25 days on 8 GPUs, Nucleotide Transformer takes 28 days across 128 GPUs, and DNABERT-2 takes 14 days on 8 GPUs for the training process[6, 10, 80]. Institutions with limited computational resources or smaller sample sizes may therefore opt to perform inference using pre-trained models, rather than training new models from scratch. Inference may be conducted through several collaborative scenarios. In one approach, a client institution transmits its input data to a model provider to obtain inference results. Alternatively, both the client and the model provider may submit their respective data and models to a mutually trusted third party for outsourced inference. These scenarios bring about privacy concerns for the client's data input, along with concerns about potential leakage of sensitive training data resulting from model sharing.

Taken together, these considerations demonstrate that collaborative studies—whether involving the integration of data across institutions or the dissemination of pre-trained models—must carefully address privacy risks to safeguard sensitive genetic and phenotypic information while supporting rigorous scienti c investigation.

2.5 Approaches to Mitigating Data Privacy Risks

Privacy enhancing techniques such as differential privacy (DP), trusted execution environment (TEE), homomorphic encryption (HE), and secure multiparty computation (MPC) offer a promising solution to privacy concerns while promoting research collaborations. DP is a mathematical framework under which calculated randomness is introduced to a dataset, ensuring that the released statistical results do not reveal whether any individual's data was included. This provides a quanti able privacy guarantee that limits the risk of re-identi cation [81]. This approach is particularly suited for releasing summary statistics, but it does not provide cryptographic security during

computation, and the injected noise can reduce accuracy, especially in the presence of outliers or skewed distributions. In contrast, TEE, HE, and MPC offer cryptographic security that protects the underlying data during computation. TEE is a hardware-based protection that refers to a secure area of the main processor that is isolated and protected from the host operating system [82, 83]. TEEs enable fast computation and provide practical advantages in terms of performance, but they remain vulnerable to side-channel attacks, which can lead to complete information leakage. HE enables computations to be performed directly on encrypted data, allowing secure analysis in cloud or server environments without exposing sensitive information [24, 84, 85, 86]. MPC enables multiple parties to collaboratively compute a function while maintaining data con dentiality, as each participant only possesses partial shares of the input [25, 26].

While differential privacy (DP) and trusted execution environments (TEE) have been explored for privacy-preserving biomedical research [81, 87], our work focuses on homomorphic encryption (HE) and secure multiparty computation (MPC) to use secure methods for collaborative association studies during computation. To contextualize the methods developed in this dissertation, we next provide a focused overview of homomorphic encryption (HE) and secure multiparty computation (MPC). This primer is deliberately scoped to highlight the theoretical concepts and practical considerations most relevant to the privacy-preserving approaches presented in subsequent chapters.

2.5.1 Homomorphic Encryption

Homomorphic encryption (HE) is well suited for tasks that require preserving data con dentiality, as it enables computations to be performed directly on encrypted data while minimizing client–server interaction. HE operates independently of inter-server communication or special hardware, which enhances security and cloud compatibility. As the input data is full encrypted prior to sharing, it provides a theoretical guarantee of privacy. However, there are two main challenges for adapting HE for machine learning inference tasks. First, HE operations are con ned to basic arithmetic operations like addition and multiplication [24]. This necessitates non-linear

operations that are common in machine learning models to be approximated into polynomial approximations. Second, HE results in larger encrypted data storage and longer computation times compared to analysis on unencrypted data, making complex model implementation often impractical and expensive. Multiple rounds of operations also result in noise accumulation, which requires another expensive step known as “bootstrapping” to manage error levels [88].

Background for Polynomials over a Ring In algebra, a ring is a set equipped with two binary operations—addition and multiplication—that satisfy properties such as associativity for both operations, distributivity of multiplication over addition, and the existence of an additive identity and additive inverses. A polynomial ring is a speci c type of ring composed of polynomials whose coef cients belong to another ring. Polynomials over a ring consist of coef cients and variables drawn from the ring, and they can be added or multiplied following conventional algebraic methods, with arithmetic operations performed according to the underlying ring structure. This algebraic framework provides the foundation for numerous mathematical disciplines, including number theory, cryptography, and algebraic geometry.

Homomorphic encryption (HE) schemes, such as Brakerski–Fan–Vercauteren (BFV) [89, 90] and Cheon–Kim–Kim–Song (CKKS) [86], rely on the algebraic structure of polynomial rings for two key reasons. First, security is guaranteed under the hardness of the Ring Learning With Errors (RLWE) problem, which generalizes the classic Learning With Errors (LWE) problem to the setting of polynomials. In RLWE, secret values and errors are represented as polynomials over a ring, and it is computationally infeasible to distinguish between noisy linear equations over this polynomial ring and uniformly random polynomials. Second, the algebraic structure of polynomials naturally supports homomorphic operations: addition and multiplication of plaintexts can be mapped to corresponding operations on ciphertexts. This property allows encrypted data to be manipulated without decryption. To perform more complex functions within HE, these functions must often be approximated as polynomials so that they remain compatible with the additive and multiplicative operations, ensuring both secure computation and correct homomorphic evaluation.

BFV is designed for exact arithmetic over integers, which makes BFV a natural choice for applications that require deterministic, exact outputs. In contrast, CKKS is optimized for approximate arithmetic, making it suitable for scenarios where slight numerical imprecision is acceptable.

The following is a brief description of a HE work ow.

1. Key Generation: Generate a public key (used for encryption), secret key (used for decryption), and evaluation keys (for homomorphic evaluation) using chosen scheme parameters.

2. Encoding: Convert input data into a polynomial representation in a nite ring. Scaling may be required depending on the scheme and use case. Data prior to encryption is referred to as `plaintext'.

3. Encryption: Encrypt the encoded polynomial plaintext into a ciphertext using the secret key.

4. Homomorphic Evaluation: Perform computations directly on encrypted data, or ciphertexts.

5. Decryption: Use the secret key to decrypt the ciphertext back to encoded plaintext, and decode back to the desired output format.

2.5.2 Secure Multiparty Computation

Secure Multiparty Computation (MPC) is a sub eld of cryptography that focuses on enabling multiple parties to jointly compute a function over their inputs while keeping those inputs private. In essence, MPC allows parties to collaborate and compute a result without revealing their individual inputs to each other [25, 26]. This is particularly useful in scenarios where privacy and con dentiality are important. There are a number of techniques MPC utilizes for security including secret sharing [91, 92], garbled circuits [25], homomorphic encryption [24], and oblivious transfer [93].

In this work, we examine secret sharing in detail within the scope of our study. In secret sharing, inputs are split into shares distributed among the parties. Computation is done on these shares, and

the nal result is reconstructed from them. Secret sharing allows private data such as genotype and phenotype values to be shared to a pre-determined number of parties such that each party cannot recover the private value without communicating with each other [91, 92]. We describe an example of secret sharing using the three-party linear replicative secret sharing based on the Ring modulo $2^{=}$ protocol as presented by Araki et al. [94]. A secret value $G_{=2^{=}}$ can be shared amongst three parties $\%_1$, $\%_2$, and $\%_3$ by rst dissecting the secret G into three uniformly-sampled random numbers G_1, G_2, G_3 where $G_1, G_2, G_3 = G$ modulo $2^{=}$. These are then distributed among computing parties such that $\%_1$'s share is (G_1, G_2), $\%_2$'s share is (G_2, G_3), and $\%_3$'s share is (G_3, G_1). Each computing party holds a partial share of the secret and therefore cannot recover the private data. We denote $[\![G]\!]$ to represent secret sharing of G. In addition, $\%_8$ shares pseudo-random number generators (PRNG) with neighboring parties: $PRNG_{8,1}$ with $\%_{8,1}$, and $PRNG_{8,1}$ with $\%_{8,1}$. These PRNGs are used to mask shared values with randomness throughout the protocol.

With this secret sharing scheme, the computing parties can jointly compute operations while keeping the data secret. Operations such as addition of two secrets can be done locally–two secrets $[\![G]\!]$ and $[\![H]\!]$ can be added locally such that $\%_1$'s share becomes $(G_1, H_1), (G_2, H_2)$, $\%_2$'s share becomes $(G_2, H_2), (G_3, H_3)$, and $\%_3$'s share becomes $(G_3, H_3), (G_1, H_1)$. This allows each computing party to own replicative shares of $[\![G + H]\!]$. Multiplication requires communication across parties to obtain the nal multiplication product, the details of which is described in section 3.2. Similar to HE, MPC requires functions to be reduced into additions and multiplications for application.

2.6 Prior Work in Privacy-Preserving Genotype-Phenotype Association Studies

2.6.1 Meta-analysis for Federated Studies

The current standard for federated genotype-phenotype studies while protecting data privacy is meta-analysis, where each data center conducts independent association study and shares summary statistics such as p-values, effect size, and standard error for aggregation [79, 95, 96, 97]. The aggregation is done with tools like METAL, where p-values and effect sizes are adjusted based on

data center sample size and standard error [98]. Meta-analysis enhances statistical power for sites with small sample sizes; however, mapping is performed locally without cross-center preprocessing, and the analysis remains constrained by the number of samples available at each data center.

2.6.2 Cryptographic Approaches

Cryptographic approaches have been applied to biological analyses through frameworks such as Flimma, a privacy-preserving differential gene expression analysis tool that combined federated learning with secure multiparty computation (MPC) [99]. Although Flimma was not designed for association studies, it represented an important step toward privacy-preserving transcriptomic analysis by exploring normalization techniques under privacy constraints. Speci cally, Flimma employed upper quartile normalization, which scaled each sample using the upper quartile of its read counts. However, this normalization approach adjusted expression levels only within individual samples and did not account for cross-sample variation. In addition, this approach is not well-suited for federated studies, where technical batch effects can arise due to variations in sequencing platforms, quality control procedures, quanti cation software, and population composition across participating institutions.

Several prior efforts explored the integration of cryptographic techniques into genotype-phenotype association studies. SecureMA provided a privacy-preserving alternative to traditional meta-analysis by encrypting aggregated statistics from each participating site [100]. PPML-Omics used differential privacy and federated learning to enable privacy-preserving, iterative training for association analysis with decentralized data [101]. Secure-GWAS, developed by Cho et al., employed secure multiparty computation (MPC) to enable joint computation for genome-wide association studies (GWAS) in a privacy-preserving manner. Building on these works, the Federated Analytics System with Multiparty Homomorphic Encryption (FAMHE) combined federated learning with multiparty homomorphic encryption (MHE) to support secure and scalable GWAS analysis [27, 102]. More recently, Secure Federated GWAS (SF-GWAS) addressed limitations of prior approaches by retaining genomic data within a federated setting, thereby avoiding the computational overhead

associated with transferring large datasets between sites [103].

Notably, these cryptography-based GWAS methods have developed secure PCA methods on horizontally-partitioned genotype data designed for federated settings. More speci cally, Secure-GWAS employs an MPC-based PCA method utilizing low rank matrix representations, while SF-PCA uses multiparty homomorphic encryption (MHE) to allow for secure PCA that is more scalable than its other secure PCA methods [27, 104]. While these methods provide guaranteed security for genotype population strati cation correction, they are computationally demanding and can require extensive time to execute. An alternative to this approach is local projection, where each site locally projects their genotype onto a loading matrix obtained from a publicly available reference panel [105] (i.e., 1000 Genomes Project [106]). This maximizes local computation without compromising accuracy, as seen by its performance in federated linear mixed models [105]. Thus, a key consideration for genotype preprocessing in federated studies is ensuring that population strati cation is adequately corrected across sites while keeping the computational burden reasonable for large-scale, horizontally partitioned datasets.

Beyond traditional association studies, prior work also explored the integration of cryptographic tools into machine learning. A key focus of these efforts was developing accurate approximations for operations that were not directly compatible with cryptographic computation. For example, Kim et al. reported an approximation of the sigmoid function used for logistic regression using polynomial approximations [107]. Gilad-Bachrach et al. introduced CryptoNet, a neural network implemented under homomorphic encryption (HE), in which the standard ReLU activation was replaced with a square function to enable secure computation [29]. Similarly, Hong et al. demonstrated HE-based inference for shallow neural networks applied to secure tumor classi cation, approximating the softmax activation to maintain privacy while performing accurate predictions [108].

2.7 Discussion of Gaps

Despite signi cant advances in genomic data analysis and privacy-preserving computation, several important gaps remain in enabling secure, collaborative genotype–phenotype association studies.

First, no existing work has focused on privacy-preserving preprocessing of gene expression data for federated studies. Historically, gene expression measurements were considered safe to share publicly, as they were not thought to uniquely identify individuals. However, recent studies have demonstrated that gene expression pro les can be exploited to recover sensitive genetic information, thereby undermining protections placed on genotype data alone [37, 38, 18]. Consequently, protecting gene expression data prior to downstream analysis has become a critical and previously overlooked challenge.

Second, while secure genome-wide association study (GWAS) frameworks have been developed, secure and federated eQTL mapping remains largely unexplored, despite its comparable biological importance. eQTL mapping introduces additional complexity compared to GWAS, as it requires analyzing a much larger phenotype space (tens of thousands of genes) and controlling for false discovery rates across an expanded number of statistical tests. These scalability challenges make it more dif cult to design ef cient privacy-preserving protocols that remain practical for large-scale studies.

Finally, as machine learning becomes increasingly integrated into genotype–phenotype association studies, new privacy risks emerge. Information leakage may occur through multiple pathways, including the exposure of model parameters. Furthermore, the rapid evolution of model architectures—such as the adoption of deep neural networks and transformers—necessitates privacy-preserving techniques that are robust to these developments. Addressing these challenges will require methods that simultaneously protect both input data and model parameters while enabling inference on complex, modern architectures.

Taken together, these gaps underscore the need for a uni ed framework that balances robust

data privacy with the ability to conduct large-scale, collaborative analyses. Future work must address privacy at every level—from genomic data to derived functional measurements—while preserving the collaborative potential necessary to advance discovery in genomics and personalized medicine.

Chapter 3: Privacy-Preserving Gene Expression Preprocessing

3.1 Overview

This aim centers around necessary gene expression preprocessing steps that are required for any standard biomedical analysis pipeline that uses gene expression data as input. Functional genomics data, such as gene expression, was once considered harmless to share. However, increasing reports of linking attacks on public phenotype data have demonstrated a critical need for enhanced privacy protection [37, 38, 18]. In this chapter, we describe the experimental setting of our methods, our design choices for privacy-preserving preprocessing methods, and validation of our methods using simulated and real-world data. We present methods to implement privacy-preserving QN, RLE normalization methods, MPC-based inverse normal transform, and local hidden covariate correction designed for privacy-preserving federated studies. Implementation details, as well as validation of our methods, can be found in our publish work [32].

3.2 Experimental Setup and MPC with Secret Sharing

The experimental setting for this study involves data partitioned across three distinct institutions, each of which independently performed gene expression quanti cation and quality control. We use MPC with secret sharing for secure federated gene expression preprocessing as it provides practical and scalable joint computation without compromising accuracy. Our protocol relies on three-party linear replicative secret sharing based on the Ring mod 2^{32} protocol as described by Araki et al. [94].

Utilizing MPC for our privacy-preserving preprocessing methods require breaking our methods down into building block of addition and multiplication (please refer to section 2.5.2 for more

details). We first describe MPC-based computation on these building blocks. While addition is trivial and can be done locally by each computing party P_i adding their local shares, multiplication relies on communication between parties. Given secretly shared $[\![G]\!]$ and $[\![H]\!]$, we can compute multiplication $[\![G \cdot H]\!]$ as in Algorithm 1, where PRNG refers to pseudo-random number generators shared between each P_i [94]. The core principle of secure multiplication in our protocol is that while the final product requires communication, the process is distributed. Each computing party can locally calculate one-third of a partial product and receive a corresponding one-third from a neighboring party. This distributed computation ensures that, through a final communication step, the parties can collectively reconstruct linear replicative shares of the multiplication product without ever revealing their individual secret shares. This multiplication can be extended to matrix operations, allowing computing parties to locally compute a portion of the multiplication product before communication. The security of addition and multiplication under the ring modulus 2^{32} with semi-honest adversaries has been proven by Araki et al. [94].

Algorithm 1 MPC Multiplication

Input. $[\![G]\!], [\![H]\!]$
Output. $[\![G \cdot H]\!]$

1: for each computing party P_i do
2: Compute $E_i = G_i \cdot H_i + G_{i+1} \cdot H_i + G_i \cdot H_{i+1}$
3: $E_i \leftarrow E_i + A_{i,i+1}$
4: $E_i \leftarrow E_i + A_{i+1,i}$
5: Send E_i to P_{i-1}
6: Receive E_{i+1} from P_{i+1}
7: end for
8: Note: $A_{i,i+1}$ is generated from $PRNG_{i,i+1}$, and $A_{i+1,i}$ is generated from $PRNG_{i+1,i}$.

3.3 Implementation Choices for Phenotype Preprocessing

A crucial step in phenotype preprocessing, especially for federated studies, is gene expression normalization. This step allows gene expression levels to be comparable across multiple studies. While there are multiple methods for normalization as described in section 2.3.2 of Chapter 2, we

provide two options for privacy-preserving normalization: QN and RLE. Both methods employ local computation and sharing aggregate summary statistics such as sum of expression values to obtain the average. TMM and RLE are more frequently used in recent studies, due to their versatile nature in dealing with RNA composition bias during normalization. We chose to implement RLE normalization in MPC instead of TMM normalization as used in other studies [43, 68], because a systematic comparison of normalization methods for bioinformatics work ows such as eQTL analysis showed similar results for these two methods [109], and RLE normalization chooses a pseudo-reference as a point of reference while TMM uses a real sample to compare library size factors. In order to protect sample privacy across all samples, we chose a method that will not reveal a real sample at any point during normalization.

Similarly, various methods exist for correcting hidden covariates in gene expression data, as outlined in subsection 2.3.2. PEER is a widely used approach in large consortium studies such as the GTEx Consortium and the Geuvadis project [43, 73]. A systematic comparison of multiple hidden covariate correction methods, including PEER and PCA, demonstrated that PCA provides the most robust performance, and it is therefore adapted in this work.

3.3.1 Validation of implementation choices

We conducted a proof-of-concept experiment to con rm our implementation choices for gene expression preprocessing methods on accuracy. We chose eQTL mapping as a comparison pipeline, and compared eGenes (a gene whose expression is signi cantly associated with a genetic variant) overlap with different combinations of gene expression preprocessing methods: (1) TMM, RLE, or QN for gene expression normalization, and (2) PCA or PEER for gene expression covariate correction (Table 3.1). We found that RLE and TMM normalization with the same phenotype correction method leads to minimal differences in eGenes discovery (2.76% difference with PCA, and 0.15% difference for PEER). We also found that using PCA or PEER as the phenotype covariate correction method results in 95.66% overlap in eGene discovery. (Table 3.1) Comparison was done in a non-secure, aggregated setting to ensure that the difference in eGenes overlap can

be attributed solely to the preprocessing method, and not secure implementation. These ndings validate our decision to implement QN and RLE as our privacy-preserving normalization methods and Principal Component Analysis (PCA) for covariate correction.

Phenotype Normalization	Phenotype Correction	# eGenes	Intersection	Overlap % (Case / Baseline)
RLE	PCA	7319	7139	100
TMM	PCA	7323	7117	97.24
QN	PCA	7801	6832	93.35
RLE	PEER	11815	7001	95.66
TMM	PEER	11823	7012	95.81
QN	PEER	11795	6999	95.63

Table 3.1: Evaluation of different phenotype preprocessing choices on eQTL mapping eGene discovery. Gene expression preprocessing choices were tested on (1) phenotype normalization–RLE, TMM, or QN, (2) phenotype covariate correction–PCA or PEER.

3.4 Privacy-preserving phenotype preprocessing methods

3.4.1 Privacy-Preserving Quantile Normalization

QN requires computing the rank of the genes based on their expression value for each sample and then replacing each rank with their across-sample rank average in a non-secure setting. Since all participating institutions have the same number of genes in their gene expression matrix, we can perform the across gene ranking for each sample locally. This ranked matrix is summed up across samples according to rank such that for M genes, there are M rank sums that are the local sum of expression values for that data owner's samples. They then secretly share the sum of gene expression values for each rank and the number of samples with the computing parties. The servers use this information to calculate the gene expression values for each rank. They do this by combining the secretly shared sums and the number of samples per site. Afterward, the servers return the average gene expression values per rank to each site. Each site can then re-order their

gene expression values based on this global rank as demonstrated in Figure 3.1A.

3.4.2 Privacy-Preserving Relative Log Normalization

RLE normalization nds a pseudo-reference representative of the samples to nd library size factors [65]. This pseudo-reference is de ned as a geometric mean across samples where = is the total number of samples (see denominator in Equation 3.1). In our privacy-preserving version, this pseudo-reference sample can be shared in plaintext amongst sites as it does not represent an actual sample. Each data owner locally calculates the log values of their gene expression values $\$_{8\,9}$;>6% for data owner : ($\$_{:}$) and secretly shares the log matrix to the three computing parties such that the computing parties have shares of the phenotype matrix, thus %-share becomes (;>6% ;>6%). The computing parties then calculate the sum of log values across the total samples = as $\sum_{E=1}^{=}$;>6% This log sum vector is revealed in plaintext to %, which then calculates the average of the log values to obtain the geometric mean across samples. This vector is designated as the pseudo-reference sample, and is returned in plaintext back to the data owners. The data owners then locally compute the log ratio between the gene expression values of their samples and the pseudo-reference. The scaling factor is then the median of these ratios for each sample, de ned as below for sample 9 and gene 8 (Algorithm 2 and Figure 3.1B).

$$\hat{B}_9 = \text{median}_8 \frac{\%_{8\,9}}{\sum_{E=1}^{=} ?_{8E}} \qquad (3.1)$$

3.4.3 Privacy-Preserving Inverse Normal Transform

After normalization, the gene expression values, which are assumed to follow a normal distribution, are transformed via inverse normal transform. Inverse normal transform refers to the process of converting values into I-scores such that the resulting data follows a standard normal distribution. For a gene expression matrix that has been normalized, each gene vector is ranked

Algorithm 2 Privacy-preserving RLE Normalization

Input. $\$_{:}$'s phenotype matrix $\%^{\$}_{89}$:
Output. per-sample size factors $\mathcal{B}$s in (3.1)

for each $\$_{:}$ do
  Compute $\log \%^{\$}_{89}$:
  Create shares of $\log \%^{\$}_{89}$ matrix and send to $\%$s
end for
for each $\%_{:}$ do
  Receive shares $\llbracket \log \%^{\$}_{89} \rrbracket$ from $\$$s
  Calculate $\sum_{E=1}^{=} \log \%_{8E}$ for = total samples and reveal to $\%_1$
end for
$\%_1$:
  1. Aggregate $\sum_{E=1}^{=} \log \%_{8E}$ from $\%_2$ and $\%_3$
  2. Return pseudo-reference vector $\frac{1}{=} \sum_{E=1}^{=} \log \%_{8E}$ to $\$$s
for each $\$_{:}$ do
  Exponentiate $\frac{1}{=} \sum_{E=1}^{=} \log \%_{8E}$ to obtain $\hat{\mathcal{J}}_1 = \prod_{E=1}^{=} ?_{8E}^{\frac{1}{=}}$
  Locally calculate $\mathcal{B} = \underset{8}{\text{median}} \frac{\%_{89}}{\hat{\mathcal{J}}_1 \prod_{E=1}^{=} ?_{8E}^{\frac{1}{=}}}$
end for

across samples, and their rank percentiles are converted into corresponding z-scores. More specifically, the rank percentiles are considered as the cumulative distribution function (CDF) values for a standard normal distribution, and they are converted to z-scores using either a look-up table or inverse normal transform procedure. Implementing inverse normal transform within a privacy-preserving framework with MPC presents two distinct challenges: (1) retrieving the rank vector, which requires comparison of secretly shared values, and (2) converting our rank vector into z-scores through a non-trivial process that often requires look-up tables even in plaintext. In our setting, look-ups are not possible since our query is kept in secret.

In order to obtain the rank vector while the expression vector is secretly shared, we adapt a secure sorting algorithm from Asharov et al. [110] that outputs a destination permutation vector. Here, the destination permutation vector refers to the permutation vector that sorts the original vector when applied. For example, destination permutation vector d; (3,4,1,2,5), sorts the original

vector ®; (3.2, 5.6, 1.7, 3.1, 6.2) when applied to it: d¹®º; (1.7, 3.1, 3.2, 5.6, 6.2). Our protocol uses this destination permutation vector as our rank vector. Brie y, the protocol takes in a bit-wise decomposed secret vector and conducts secure radix sort starting from the least signi cant bit. Instead of generating rank percentiles from the permutation vector, we use the permutation vector outputted from this protocol to reorder a given set of ordered I-scores.

Our MPC-based inverse normal transform protocol takes in the resulting secretly shared permutation vector ÈdÉ from secure sorting method, a secretly shared identity permutation È É, and secretly shared ordered I-scores ÈIÉ as input. With a known number of total samples #, ordered I-scores can be viewed as a set of # values that can be designated prior to the protocol. The output of our protocol is a secretly shared set of re-ordered I-scores. We rst apply ÈdÉ onto È É to obtain Èd^1 É. Èd^1 É can be applied to the ordered I-scores ÈIÉ to reorder the I-scores to their data-corresponding position. We use the APPLYPERM protocol detailed in Asharov et al. [110] to apply secretly shared permutations on secretly shared vectors. This requires masking both ÈIÉ and Èd^1 É by a random secretly shared permutation ÈcÉ, and revealing the randomly masked vector È$c^1 d^1$ ºÉ to apply on Èc^1IºÉ. (Figure 3.1C) Our secure inverse normal transform protocol requires the rank vector ÈdÉ as its sole private input. The identity vector and the ordered z-score vector I are also necessary for the protocol; however, since these can be generated from the total number of samples across institutions, they are not considered private information and are provided directly by the data owner.

Figure 3.1: Privacy-preserving phenotype preprocessing methods. (A) QN–data owners their aggregate sums to one of the computing servers. (B) RLE–log expression matrices are secretly shared, then aggregated to reveal the pseudo-reference. Per-sample size factor calculation is done locally. (C) Inverse Normal Transform–the identity permutation is secretly shared to compute the inverse permutation, which is then applied to the ordered z scores. Double brackets indicate secret sharing.

3.4.4 Privacy-Preserving Phenotype Covariate Correction

In a non-secure setting, normalized and inverse normal-transformed gene expression matrix is corrected for hidden covariates using PEER or PCA. For this study, principal components from PCA are used instead of PEER factors for gene expression covariate correction as previous studies report that it is more scalable, interpretable, and robust compared to PEER factors [111]. After privacy-preserving and federated normalization of the user's choice and inverse normal transform, the normal transformed gene expression values are returned to the participating sites. They then locally compute hidden variable correction with PCA [111] with their own samples and residualize their gene expression values based on their local principal components. The fully processed gene expression matrix is then ready to be secretly shared to computing parties for downstream pipeline.

3.5 Simulated Study

To determine how different preprocessing choices in uence the overall accuracy and sensitivity of a bioinformatics pipeline, we conducted a federated simulation study. We used eQTL mapping as a representative example to assess the accuracy of our preprocessing methods. Our simulation draws upon the methods outlined by Zhou et al. [111]. Additionally, we divided the data randomly into three distinct sites and mimic batch effects. We repeated this for two different splitting scenarios: one had 300, 250, and 120 samples across three different sites (denoted as 300/250/120), and the other had 300, 300, and 70 samples (denoted as 300/300/70) across three different sites. The site information is simulated via addition of noise from different distributions and was incorporated as a covariate in our correction for hidden variables. The goal is to assess whether different gene expression pre-processing steps are able to remove covariates and nd the correct SNP as the effect SNP (i.e., an eVariant). Thus, the gene expression matrix was modeled as a linear combination of genotype matrix, binary effect indicator matrix (indicates whether a SNP is an eVariant for a given eGene), hidden and known covariates (continuous hidden covariates modeled from normal distribution and known covariates include site information and sex), effect size matrices, and noise.

More speci cally, we chose 1000 random genes from the GTEx whole blood data and chose 1000 SNPs that are closest in distance to each gene. For ==670 samples, ?=1000 genes, @=1000 local SNPs per gene and : covariates, we simulated the expression matrix . as linear combination of the genotype matrix , covariate matrix -, and error as in Zhou et al. (Equation 3.2)

$$= \dot{} ? = = @ ? \quad @ ? \quad @ ^{1} ? \; , \; = - : \quad : ^{2} ? \; , \; = ? \tag{3.2}$$

is denoted as gene-wise multiplication. Each component of the simulation is detailed below.

- is a 670 (samples) 1000 (SNPs) 1000 (genes) matrix created from real GTEx whole blood data, where 1000 genes were taken from the phenotype matrix, and genotypes of the 1000 closest SNPs were chosen based on distance from gene position.
- is a binary effect indicator matrix modeled from 8=><¹@– ?A>1 = 1•@º for each gene.
- $_1$ is an effect size matrix for , drawn from #¹0– 1º.
- - is a covariate matrix consisted of both known (including site information) and unknown covariates.
- $_2$ is an effect size matrix for -, drawn from #¹0– 1º.
- is an added error matrix, drawn from # with different noise parameters according to site.

The resulting expression matrix . was corrected according to (1) Normalization: QN or RLE, (2) aggregated or federated covariate correction method with PCA. The different combination of strategies is described in Table 3.2. The objective of each experiment was to assess how well each preprocessing method accurately corrects the phenotype matrix and ultimately identi es effect SNPs as compared with the effect indicator matrix $_{2>A}$ which is corrected for linkage disequilibrium such that any SNP that has high correlation (j'j ¡ 0•9) with any of effect SNPs in are marked as 1. We compared the p-values from eQTL mapping with $_{2>A}$ That is, if the p-value

obtained from association is small, then the probability of the SNP to be an eVariant is high. This allows us to model the task as a binary classi cation, thus allowing us to calculate area under the precision-recall curve (AUPRC) and area under the receiver-operating characteristic curve (AUROC) of each strategy compared to the ideal scenario, where all covariates are known. Simulation was repeated 100 times with different parameters for number of covariates, proportion of variance explained by genotype, and proportion of variance explained by covariates. The reported AUPRC and AUROC per pre-processing method is the average of those 100 simulations. In a real-world scenario, the complete set of covariates affecting gene expression values is unknown. To isolate the true effect of each preprocessing strategy on eQTL mapping, we performed the analysis without implementing the secret-sharing scheme, thus the accuracy loss can be solely attributed to the choice of strategy for preprocessing.

Label	Description
Aggregated QN	Gene expression data is aggregated in one of the sites. QN normalization and covariate correction via PCA is performed on the aggregated data.
Aggregated RLE	Gene expression data is aggregated in one of the sites. RLE normalization and covariate correction via PCA is performed on the aggregated data.
Federated QN	Gene expression data is split across sites. Federated QN is performed, and local PCA covariate correction is performed for each site.
Federated RLE	Gene expression data is split across sites. Federated RLE is performed, and local PCA covariate correction is performed for each site.
Unadjusted	Gene expression data is aggregated in one of the sites. No normalization or hidden variable correction is performed.

Table 3.2: Simulation strategies for phenotype pre-processing comparison. eQTL p-values were compared with the ground truth effect indicator matrix to evaluate area under the receiver operating characteristic curve (AUROC) and precision-recall curve (AUPRC).

Aggregated gene expression preprocessing produced the highest relative AUPRC and AUROC (0•998 and 0•904 for QN, and 0•998 and 0•908 for with RLE, respectively, Figure 3.2A). This strategy represents a gene expression data sharing scheme, where each site's data is shared publicly with a participating site to perform the normalization and hidden variable correction on aggregated

data. The methods with the next highest relative AUPRC and AUROC were the federated preprocessing strategy, with different normalization methods (0•995 and 0•869 for QN, and 0•996 and 0•873 for RLE). Note that we did not nd signi cant differences between RLE and QN normalizations in both aggregated and federated versions. We con rmed these trends with 300/300/70 data split, where both aggregated and federated corrections outperformed the Unadjusted case (Figure 3.2B). While the AUROC was not affected by different strategies, there was an expected small drop in AUPRC in settings where gene expression data was split. We note that the PR curve is more informative for this study due to the prevalence of class imbalance. The ROC curve can be misleading because it uses the False Positive Rate (FPR), which is calculated with the number of negative samples. On the other hand, the PR curve speci cally measures the balance between precision (true positive rate) and recall (false discovery rate), providing a more accurate evaluation of performance on the minority class.

Overall, these results suggest that either of the two privacy-preserving phenotype preprocessing options: federated QN with local PCA or federated RLE with local PCA show better performance than uncorrected raw data, despite its minimal drop in accuracy compared to the aggregated setting.

Figure 3.2: Effect of gene expression pre-processing on QTL mapping performance. (A) Data is split to even number of samples per site. Aggregated data sharing outperforms all other strategies, followed by federated preprocessing methods. (B) Data is splitted to uneven number of samples per site. This does not affect the performance.

3.6 Real-world data

We also demonstrated the accuracy of our methods when there is a batch effect present in the splits using real-world data. For that, we applied it to datasets collected from three different studies: GTEx project, Geuvadis project, and Taylor et al. 2024. [43, 73, 112]. We obtained both the transcripts per million (TPM) and read count data for 670 whole blood samples from the GTEx public data portal (GTEx Portal) [43]. We processed the gene expression data according to the pipeline outlined by the GTEx consortium [43]. This involved retaining the expression data based on the following thresholds: TPM 0•1 in at least 20% of samples, read counts 6 in at least 20% of samples. To mimic a multi-site gene expression study, we split the samples into three sites using two split ratios: (300,300,70) and (300,250,120).

The Geuvadis project [73] utilized lymphoblastoid cell lines (LCL) from 462 individuals of the

1000 Genomes Project[106]. From their publicly available data deposit [113], we acquired read count data encompassing 58,813 genes across 462 samples. After aligning these data with the 1000 Genomes Project genotype data and excluding samples used in the Taylor et al., 2024 study, we selected a subset of 150 samples. Gene expression data was ltered based on the following criteria: TPM 0•1 in at least 20% of the samples and read counts 6 in at least 20% of the samples.

Similarly, Taylor et al., 2024 [112] used LCLs from 731 individuals from the 1000 Genomes Project. We obtained their ltered read count data, which included 20,154 genes and 734 samples. We selected 300 samples that have matching 1000 Genomes genotype data and are different than the Geuvadis samples we selected. The gene expression data for this subset were ltered based on the same thresholds: TPM 0•1 in at least 20% of samples and read counts 6 in at least 20% of samples.

The gene expression data was not only collected in different laboratories but there were also differences in how the expression values were quanti ed. Both GTEx and Geuvadis projects used RNA-SeQC v1.1.9[114] and RSEM v1.3.0[115] while Taylor et al., 2024 used Salmon v1.5.2[116] for the quanti cation. Most importantly, data from each data source represents an uneven population distribution. The GTEx project contains data from primarily European and a small number of African American and Asian populations, the Geuvadis project contains mostly European samples and a small number of African American samples, while Taylor et al., 2024 contains samples from African American, European, East Asian, Americas, and South Asian populations.

To validate our approach, we employed Principal Component Analysis (PCA) to visualize the impact of three preprocessing strategies on batch effects by plotting their rst two principal components. The strategies included: (1) Unadjusted, representing a simple aggregation of the three datasets without any processing; (2) Aggregated PCA (denoted as privateQTL-I), which serves as a centralized baseline by performing normalization and PCA-based correction on the combined data; and (3) our proposed Separate PCA (denoted as privateQTL-II), which applies federated normalization and local PCA using our privacy-preserving protocols.

The initial unadjusted data clearly demonstrated the presence of batch effects, evidenced by

the distinct clustering of data points based on their data source. The subsequent application of both the aggregated and our privacy-preserving preprocessing methods successfully corrected these effects, as visualized by the intermixing of the data clusters. Crucially, the results from our privacy-preserving phenotype preprocessing method in Figure 3.3C show a corrective effect identical to that of the aggregated, non-private baseline. Ultimately, this validates our approach as a robust solution for correcting for hidden covariates in the presence of data heterogeneity.

Figure 3.3: Batch effects before and after each covariate correction method. We performed PCA on (A) raw gene expression values, (B) PCA on gene expression values after centralized normalization and hidden covariate correction, and (C) PCA on gene expression values after federated RLE normalization and per-site hidden covariate correction via PCA. We found that both strategies resulted in very similar results.

3.7 Discussion

In this aim, we developed a set of privacy-preserving methods for phenotype preprocessing that integrate seamlessly into standard bioinformatics pipelines. We give the user option between two widely used methods: QN and RLE for privacy-preserving normalization. Both versions maximize local computation and shares only aggregate information such as total sample size and sum of values. This allows for maximized computation ef ciency while guaranteeing accuracy and privacy. Such computational ef ciency is crucial for cryptographic techniques such as MPC where there

is an inherent communication overhead required from the computing parties. Other commonly used normalization methods such as TMM are not viable for privacy-preserving implementation as they reveal information about private data input. We have demonstrated that our federated methods show competitive accuracy with the aggregated, non-secure counterpart from the overlap of eGenes from eQTL mapping.

We propose a novel MPC-based method for secure inverse normal transform, where normalized gene expression vectors are ranked and converted into z-scores by applying secure permutation vectors. This is the rst secure implementation of inverse normal transform to our knowledge, as the formula is non-trivial even in a non-secure setting. Instead of using a lookup procedure where both the input and output are secretly shared, our procedure considers both the input rank vector and the output z-scores as a rearrangement of a set of values according to a given permutation. This allows the computing parties to `rearrange' the values without ever knowing the data. As our inverse normal transform protocol is gene-vector based, it allows for easy parallelization.

We employed a local PCA method for gene expression hidden covariate correction method. While secure and federated PCA techniques are available, they are computationally demanding and can require extensive time to execute [27, 104]. We hypothesized that implementing a cross-institutional normalization method in conjunction with local hidden covariate correction would effectively mitigate the bulk of data heterogeneity. Through our analysis of both real and simulated datasets, we demonstrated that our method yields preprocessing results that are comparable to those obtained from an aggregated, centralized approach.

Our privacy-preserving gene expression preprocessing methods, which utilize local computation and Multi-Party Computation (MPC), are modular in design. The user has the option to use the methods altogether in a work ow or separately on their own, allowing for exible adaptation to other phenotype preprocessing pipelines.

We also assessed our methods comprehensively in two distinct scenarios for eQTL mapping. First, we used a simulated dataset, which provided a controlled environment with a known ground truth. This enabled us to precisely quantify performance metrics. Second, we validated our meth-

ods using a multi-center dataset to account for data heterogeneity inherent within real-world data. While ground truth is unknown for real-world eQTL data, we were able to visualize how our method corrects for hidden covariates as compared to aggregated, `gold standard' approach and unadjusted approach.

In conclusion, this work successfully addresses the critical need for privacy-preserving solutions in bioinformatics by developing a modular suite of methods for phenotype preprocessing. We have demonstrated that our approach, which leverages a combination of local computation and Multi-Party Computation (MPC), offers a computationally ef cient, accurate, and viable alternative to traditional centralized pipelines. By providing a practical framework that balances the demands of scienti c discovery with the imperative of data privacy, our research not only validates the utility of secure computation in genomics but also lays the groundwork for future federated and privacy-preserving studies.

Chapter 4: Scalable Pipeline for Secure and Federated eQTL mapping

4.1 Overview

This chapter describes practical ways to apply cryptographic techniques for privacy-preserving eQTL mapping pipeline designed for federated studies. Building on the non-secure eQTL mapping pipeline described in section 2.3.3 of Chapter 2, we present privateQTL, a MPC-based tool for secure and federated cis-eQTL mapping [32]. PrivateQTL is an end-to-end protocol that includes federated and privacy-preserving (a) regression for both nominal (nds associations) and permutation runs (generates null distribution for cis-associations), (b) post-processing for multiple hypothesis testing correction. Details on our secure implementation and benchmarking can additionally be found in our publication [32].

We offer two variations of privateQTL, each designed to meet speci c requirements. One version is compatible with existing data-sharing protocols, while the other is forward-looking and enhances privacy protections to guard against re-identi cation through linking attacks on gene expression data. PrivateQTL-I thus allows gene expression preprocessing on aggregated, public data, while privateQTL-II employs federated, privacy-preserving preprocessing methods described in Chapter 3 (Figure 4.1). Both versions of privateQTL keep genotype data as private. We describe detailed methodologies of each step below.

Figure 4.1: Schematic overview of privateQTL. (A) In both privateQTL-I and II, the genotype preprocessing is done using a local projection method for population strati cation. This allows sites to project their data onto the same reference PCA space, while keeping the data local. In privateQTL-I, gene expression data is shared across sites and pre-processing is done on the aggregated data without privacy-preservation. In privateQTL-II, gene expression data is shared via secret sharing and normalization takes place with MPC. The normalized gene expression data are then sent back to the sites for covariate correction. (B) In both privateQTL-I and II, eQTL mapping and null distribution generation are done using MPC via secret sharing of both gene expression and genotype data. Double bracket indicates secretly shared data.

4.2 Privacy-Preserving Genotype Population Strati cation Correction

In the beginning of the protocol, each data client sends its set of gene and SNP IDs to a designated computing party, which calculates and shares the intersection with its clients. This functionality also includes automatic ltering of each client's data based on the common genes and SNPs identi ed. It is important to note that gene and SNP names or IDs do not contain private information.

Prior to the mapping step, the genotype requires preprocessing due to population strati cation, which is typically addressed in a non-secure setting through PCA followed by residualization. Implementing strati cation correction in federated studies securely requires the use of PCA on horizontally partitioned data. While methodologies for this exist, they often demand extensive time and computational resources [27, 104]. As such, we adapt a projection method where data centers locally project their genotypes onto a reference panel consisting of principal components that contains their population subgroup [105] (Figure 4.1A). This is done after each site performs their own quality control on their genotype data. The hypothesis is that if the sample population is a subset of the reference population, the reference panel would represent population structure that encompasses sample genotypes. This projection-based correction method allows for each site to correct and residualize their data locally.

Similar to phenotype preprocessing, we conducted a proof-of-concept experiment to con rm accuracy on genotype projection method. Please refer to subsection 2.3.3 for details on eQTL mapping and terminology. We compared eGenes overlap with different combinations of genotype projection and gene expression preprocessing methods. We used GTEx whole blood samples[43] and split them into three sites, and conducted genotype preprocessing in an aggregated setting (denoted as Agg PCA) or local setting with projection onto 1000 Genomes Project (denoted as Projection) [106]. Results show that for the same set of phenotype preprocessing methods (details of phenotype preprocessing methods and implementation results outlined in subsection 2.3.2 of Chapter 2 and Chapter 3 respectively), projection-based genotype correction recovers around

95% of eGenes compared to aggregated PCA, where eGenes are de ned as genes that have significant correlations with SNPs (Table 4.1). Even with different phenotype combinations, the drop in eGene discovery overlap does not exceed 5% compared to aggregated genotype correction. Our results demonstrate that our projection-based genotype correction is a viable and ef cient alternative to computationally demanding federated PCA methods, effectively mitigating population strati cation with a minimal impact on eGene discovery.

Genotype Correction	Phenotype Normalization	Phenotype Correction	# eGenes	Intersection	Overlap % (Case / Baseline)
Agg PCA	RLE	PCA	7319	7139	100
Projection	RLE	PCA	7130	6922	94.58
Agg PCA	TMM	PCA	7323	7117	97.24
Projection	TMM	PCA	7108	6846	93.54
Agg PCA	QN	PCA	7801	6832	93.35
Projection	QN	PCA	7686	6738	92.06
Agg PCA	RLE	PEER	11815	7001	95.66
Projection	RLE	PEER	11786	6982	95.40
Agg PCA	TMM	PEER	11823	7012	95.81
Projection	TMM	PEER	11889	7011	95.79
Agg PCA	QN	PEER	11795	6999	95.63
Projection	QN	PEER	11812	6985	95.44

Table 4.1: Evaluation of genotype preprocessing choices on eQTL mapping eGene discovery. Preprocessing choices were tested on (1) genotype covariate correction—aggregated PCA or projection onto reference, (2) phenotype normalization—RLE, TMM, or QN, (3) phenotype covariate correction—PCA or PEER.

4.3 privateQTL: Secure and Federated cis-eQTL mapping

As described in the section 2.3.3 in Chapter 2, the phenotype space expands from a single vector in GWAS to a matrix in eQTL studies, making individual testing of each SNP–gene pair computationally impractical. Matrix eQTL introduced association computation via matrix multiplication [74], which was accelerated in fastQTL through beta distribution-based null distributions, reducing required permutations from 10–000 to 1–000 [75], and further optimized in tensorQTL using tensor operations [76]. privateQTL is modeled after tensorQTL, enabling ef cient, privacy-preserving cis-eQTL mapping in federated settings.

4.3.1 Center and normalizing for cis-eQTL mapping

As our protocol is designed for cis-eQTL mapping, where the genotype matrix varies based on gene position, privateQTL maps eQTLs one gene at a time. It is parallelized, allowing simultaneous mapping of multiple genes. For that, each site locally lters SNPs to create a cis-window genotype matrix, which contains SNPs only within 1MB cis-window of the gene. After fully preprocessing the cis-window genotype matrix and gene expression vector, the data owners center and normalize them to have zero mean and unit sum of squares before being secretly shared to the three computing servers. This is to ensure that the correlation matrix can be obtained from matrix multiplication, as detailed in MatrixQTL [74]. This center and normalize step requires the mean and variance of each matrix to be shared publicly to the data owners. For the inverse normal transformed phenotype vector, the mean and variance are already known, as the expression values were replaced by I-scores. The data owners ($s) can locally normalize gene expression values as the following: . For the genotype matrix, each $ sums up the genotype values ( ) for their local samples (") per variant and shares to aggregates the sum values and returns the total aggregate sum back to the $s, who then calculate the mean. In a similar manner, each $ aggregates their local sum of , sends to , and receive the total

aggregate $\left(\sum_{9=1}^{\#} {}^{1}{}_{9} \grave{}_{6}\right)^{2}$. The $s locally calculate f_6 from this aggregate, and ultimately calculate $_9 = \frac{\rho^{9} \grave{}_{6}}{f_6^2}$ for each variant. These fully pre-processed genotype matrix and gene expression vector are secretly shared to the computing parties.

4.3.2 One-shot matrix multiplication for nominal and permutation runs

Due to the vast number of association tests performed for SNP-gene pairs, eQTL mapping requires permutation modeling to control false discovery rates. In a non-secure eQTL mapping setting, the fully preprocessed genotype and phenotype matrices are multiplied once to obtain a nominal correlation matrix. Then the phenotype is randomly shuf ed around 1000 times, and multiplied each time with the genotype matrix to form a null distribution for false discovery control [75]. This permutation modeling is non-trivial in a secure setting, as it requires repeated MPC-multiplication. Instead, the process can be expedited for cis-eQTL mapping where the gene vector can be shuf ed 1000 times and concatenated to form a phenotype matrix.

Upon receiving the secretly shared, fully-processed cis-genotype matrix and phenotype vector, the computing parties rst shuf e the gene vector for 1000 times with secretly shared permutation $\langle c \rangle$, where $c = c_1 \quad c_2 \quad c_3$ and represents composing permutations where one permutation is applied on another. We denote $\langle c \rangle$ as secretly shared permutation c where holds $\%\ {}^{1}c_8 - c_{8,1}{}^{o}$. During each permutation iteration, there are three permutations c_1, c_2, and c_3 that are shared between the computing parties in a linear replicative manner. The shuf ing protocol was adapted from Asharov et al. [110], where two of the three computing parties iteratively apply their permutation and reshare to the third computing party, after masking with randomness generated by their shared pseudorandom number generator (PRNG). The resulting gene expression matrix should have 1001 columns, where the rst column represents the original gene expression vector and the rest are the shuf ed gene expression vectors to be used in null distribution (Figure 4.2). The computing parties compute matrix multiplication of these two secretly shared matrices to obtain the correlation matrix $' = \%\,)$, using the MPC-multiplication protocol described in Algorithm 1.

After we perform the above calculation for all genes, the resulting correlation matrix ' (genotypes by gene expression) is revealed to one of the computing servers for downstream statistical analysis for p-value adjustment and FDR control. We use the permutation run to model a beta distribution of p-values, which is then used to correct the nominal p-values. We compare the most signi cant variant from the nominal run with those from the permutation run to adjust the nominal p-value for the most signi cant variant. This adjusted p-value for each gene is used to calculate q-values across genes, and these q-values are used for eGenes cutoff (0•05).

Figure 4.2: Implementation details of privateQTL mapping. During the mapping step, the secretly shared phenotype vector is shuf ed 1000 times with a secretly shared random permutation to create a phenotype matrix. The matrix multiplication result will include both the nominal and permutation runs. Double bracket indicates secretly shared data.

4.4 Dataset description

PrivateQTL was validated by comparing accuracy and scalability with different privacy-settings: (1) non-secure GTEx pipeline in an aggregated setting, (2) both versions of privateQTL, and (3)

meta-analysis where the data remains siloed and summary statistics are aggregated. For the validation dataset, we rst demonstrated the accuracy of our methods using GTEx whole blood data from 670 individuals, which we partitioned across three centers in a 300/250/120 split. We then validated our methods' ability to correct for batch effects by applying them to a multi-center dataset composed of data from the GTEx project, the Geuvadis project, and the Taylor et al. 2024 study [43, 73, 112]. Phenotype collection was identical to the process described in Aim 1; details on genotype collection and quality control are provided below.

For GTEx data, the controlled-access post-QC genotype was obtained from dbGAP with 46,569,704 total variants (accession code: phs000424). Genotype data was ltered to retain bi-allelic and autosomal SNPs. We only retained blood samples that have matched gene expression data (670 samples). Additionally, we ltered the genotypes with missingness of 0.15, variant quality 15, and minor allele frequency of 5%.

For the genotype from Geuvadis and Taylor et al. data, we matched 5–751–798 SNPs from the 1000 Genomes Project with GTEx, selecting only autosomal and biallelic SNPs. Duplicated SNPs in the vcf were removed from the GTEx, Geuvadis, and Taylor et al. 2024 genotype data to ensure consistency across datasets.

4.5 Benchmarking privateQTL with other eQTL mapping techniques

We used tensorQTL on aggregated data to benchmark our tool, where genotype PCA was done without projection onto reference and phenotype was RLE normalized and corrected for covariates with PCA (mimicking the GTEx pipeline). We chose RLE normalization instead of the TMM normalization used by the GTEx pipeline since RLE and TMM have similar performance for eQTL mapping and RLE does not require an actual sample as a reference [69].

We set our privacy-preserving comparator as meta-analysis, where each data owner conducts their own eQTL mapping study and attempts to aggregate summary statistics afterwards. In meta-analysis, each genotype data was projected onto reference genome and residualized, and the phe-

notype data was locally RLE normalized and residualized with their own principal components. We assumed each eQTL mapping study used tensorQTL for their eQTL mapping method. For summary statistic aggregation, we used inverse variance based METAL with xed effect modeling [98]. For multiple hypothesis testing correction in meta-analysis, we adapted the method used in Zeng et al.[79], which uses Bonferroni correction for variant-level false discovery control, since permutation-based modeling cannot be used when data should not be shared across sites. It then uses Benjamini-Hochberg method [117] for false discovery control for gene-level. Genes with Benjamini-Hochberg adjusted p-values less than 0•05 were deemed as eGenes, and variants with Bonferroni adjusted p-values less than 0•05 were deemed as eVariants.

We also compared privateQTL results with another widely used non-secure mapping method LIMIX, that employs a univariate association test within a linear mixed model framework [59]. In our study, we utilized LIMIX (v.2.0.x) st_scan function with the same 1MB cis-window as in other methods used in this work. We used eigenMT [118] to correct for gene-level multiple hypothesis testing and Storey's q-values[119] for variant-level false discovery control.

4.6 privateQTL Delivers Accurate Results with Federated GTEx Data

Using GTEx whole blood data, we simulated a federated setting by splitting 670 samples into three centers using a 300/250/120 split. We compared both versions of privateQTL with GTEx pipeline in an aggregated setting, and meta-analysis in federated setting. We evaluated privateQTL in a simulated multi-center setup, where a single data owner—simulated as one CPU core—secretly shared data with three computing parties, each simulated on a separate CPU core with network communication.

4.6.1 eGenes Discovery Analysis

We rst aimed to evaluate whether our tool can identify the same set of genes (eGenes) that exhibit signi cant associations between expression levels and SNPs (referred to as eVariants). To do

this, we compared the overlap of eGenes detected by privateQTL-I, privateQTL-II, meta-analysis, and the GTEx pipeline, using a 5% false discovery rate (FDR) threshold. Overall, privateQTL-I eGenes overlapped with 93.2% of GTEx pipeline eGenes, privateQTL-II eGenes overlapped with 91.3% of GTEx pipeline eGenes, while meta-analysis eGenes overlapped with only 76.1% of GTEx pipeline eGenes (Figure 4.3A,B). The effect sizes for the most signi cant eGene-eVariant pairs were identical to those obtained with the GTEx pipeline effect sizes for the same eGene-eVariant pair for all methods (Figure 4.3C-E). Note that some of the most signi cant eVariants found by both PrivateQTL versions and meta-analysis did not match the most signi cant eVariants found by GTEx pipeline. Therefore, we asked whether these non-overlapping most signi cant eVariants from different methods are in high Linkage Disequilibrium (LD) with each other. The expectation was that while each method might not identify the exact same variant as the most signi cant, it could nd a variant in high LD with the one found by the GTEx pipeline. We found that most signi cant eVariants that are produced by both privateQTL versions and meta-analysis are in high LD with the most signi cant eVariant of the same eGene from the GTEx pipeline (Figure 4.3F-H). As a sanity check, we also found that most signi cant eVariants are in close proximity to the Transcription Start Site (TSS) of the corresponding eGene for meta-analysis and both versions of privateQTL (Figure 4.4). Overall, meta-analysis had less power in nding eGenes, while both versions of privateQTL had a higher overlap with the GTEx pipeline.