nature genetics

Letter

https://doi.org/10.1038/s41588-025-02109-1

Secure and federated genome-wide association studies for biobank-scale datasets

Received: 29 November 2022

Accepted: 28 January 2025

Published online: 24 February 2025

Check for updates

Hyunghoon Cho[1,2,3,8] ✉, David Froelicher[3,4,8], Jeffrey Chen[3,4], Manaswitha Edupalli[3], Apostolos Pyrgelis[5], Juan R. Troncoso-Pastoriza[6], Jean-Pierre Hubaux[5,6] ✉ & Bonnie Berger[3,4,7] ✉

Sharing data across institutions for genome-wide association studies (GWAS) would enhance the discovery of genetic variation linked to health and disease[1,2]. However, existing data-sharing regulations limit the scope of such collaborations[3]. Although cryptographic tools for secure computation promise to enable collaborative analysis with formal privacy guarantees, existing approaches either are computationally impractical or do not implement current state-of-the-art methods[4–6]. We introduce secure federated genome-wide association studies (SF-GWAS), a combination of secure computation frameworks and distributed algorithms that empowers efficient and accurate GWAS on private data held by multiple entities while ensuring data confidentiality. SF-GWAS supports widely used GWAS pipelines based on principal-component analysis or linear mixed models. We demonstrate the accuracy and practical runtimes of SF-GWAS on five datasets, including a UK Biobank cohort of 410,000 individuals, showcasing an order-of-magnitude improvement in runtime compared to previous methods. Our work enables secure collaborative genomic studies at unprecedented scale.

Secure computation frameworks from modern cryptography offer promising strategies to address privacy concerns in collaborative genomic studies[3,7–9]. These techniques allow a group of parties to jointly analyze their data by exchanging encrypted information while ensuring that each party's data remain private from others. Although recent work has illustrated the potential of this strategy for genome-wide association studies (GWAS)[4–6], existing methods remain limited in practical utility. Prior work based on the framework of secure multiparty computation (MPC)[4] is prohibitively slow for large biobank-scale cohorts (as demonstrated in this work) and requires the external sharing of entire input datasets in encrypted form, which may not be feasible in practice due to security risks. More recent methods based on homomorphic encryption (HE) schemes[5,6] improve efficiency but fail to support the standard analysis pipelines commonly used by researchers due to their computational complexity. These pipelines include those based on principal-component analysis (PCA) and linear mixed models (LMMs), which provide strategies to account for population structure within a study cohort to accurately estimate association signals[10,11].

We developed secure federated genome-wide association studies (SF-GWAS), a secure and federated algorithm for multisite GWAS,

[1]Department of Biomedical Informatics and Data Science, Yale School of Medicine, New Haven, CT, USA. [2]Department of Computer Science, Yale University, New Haven, CT, USA. [3]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [4]Computer Science and AI Laboratory, MIT, Cambridge, MA, USA. [5]School of Computer and Communication Sciences, EPFL, Lausanne, Switzerland. [6]Tune Insight SA, Lausanne, Switzerland. [7]Department of Mathematics, MIT, Cambridge, MA, USA. [8]These authors contributed equally: Hyunghoon Cho, David Froelicher. ✉e-mail: hoon.cho@yale.edu; jean-pierre.hubaux@epfl.ch; bab@mit.edu

https://doi.org/10.1038/s41588-025-02109-1

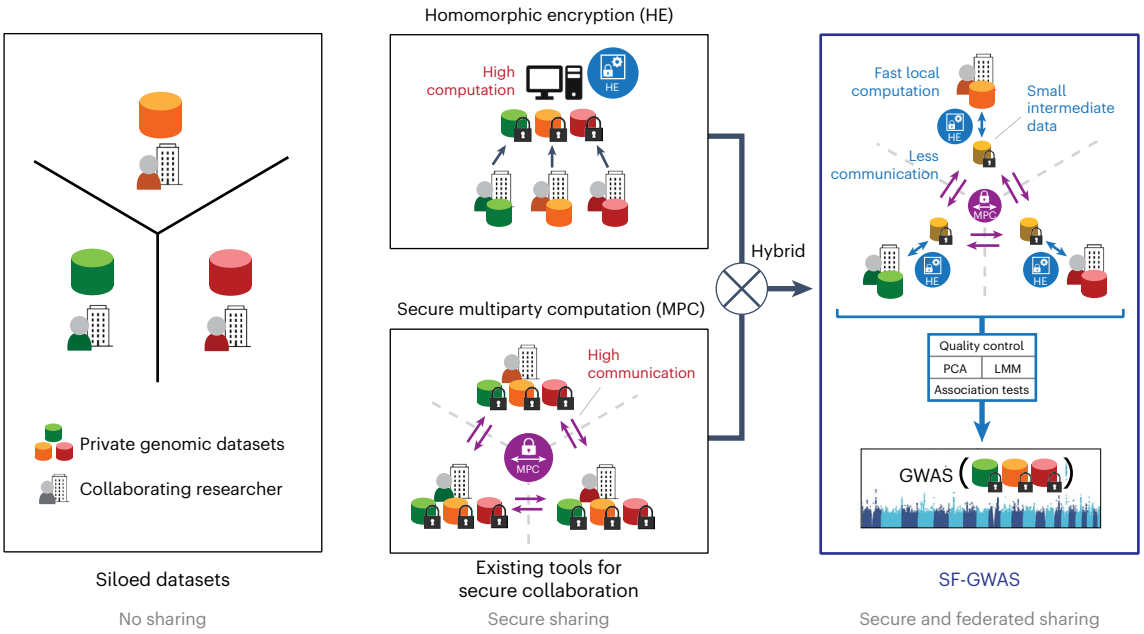


Fig. 1 | Overview of SF-GWAS. SF-GWAS addresses a common challenge faced by collaborating researchers at different institutions who wish to conduct a joint study but are unable to do so due to data-sharing limitations based on privacy concerns. Modern cryptographic solutions for jointly analyzing private datasets with formal privacy guarantees include HE and secure MPC. However, existing solutions for GWAS have limited scalability due to the high costs of computation and communication incurred by these frameworks. SF-GWAS is built upon a combination of HE and MPC to enable secure and federated computation, where large private datasets are locally kept by each data holder and only small intermediate data are encrypted and shared among the collaborators to carry out complex global computations. SF-GWAS introduces practical, secure and federated algorithms to support two essential workflows for GWAS based on PCA and LMMs. The final result includes GWAS association statistics, jointly computed over all private datasets while preserving data privacy.

to enable collaborative genomic studies at scale with cryptographic privacy guarantees (Fig. 1). SF-GWAS builds upon the following two key conceptual advances. First is our 'federated' approach to secure computation, whereby each input genomic dataset is kept at the respective source site. This approach minimizes computational costs by both avoiding large data transfers between sites and allowing the use of efficient cryptographic operations that leverage the unencrypted input data locally available at each site. To enable this strategy, we combine MPC and HE—two cryptographic schemes for analyzing encrypted (masked) private data, which previous work on secure GWAS has separately leveraged—into a hybrid framework (Methods). Our framework employs HE for local computations over large matrices and vectors (for example, multiplication between a genotype matrix and a vector of individual phenotypes) and MPC for nonlinear operations such as division and sign functions (for example, used in statistical calculations) to improve numerical precision and efficiency. Previous work on a multiparty extension of HE[5] did not utilize computational MPC routines, which we found crucial for our large-scale analysis pipelines due to the wide range of values encountered during the analysis.

Secondly, SF-GWAS introduces an efficient algorithmic design to support the federated execution of various end-to-end GWAS pipelines. Our workflow includes, in addition to association tests, quality control (QC) procedures and two standard strategies to account for population structure and sample relatedness in the cohort[10,11]: PCA or LMMs. Both PCA and LMMs involve highly complex linear algebra computations, which thus far have limited the development of privacy-preserving algorithms for these tasks that can be applied to large biobank datasets. Using a set of algorithmic strategies aimed at optimizing performance in a distributed setting, such as techniques to maximize the use of local plaintext (unencrypted) data, we developed practical secure algorithms for both PCA- and LMM-based GWAS workflows, demonstrating the broad utility of our framework (Methods and Supplementary Note).

We first compared SF-GWAS with the prior state-of-the-art method for secure PCA-based GWAS[4], referred to as S-GWAS. We used both methods to analyze the three datasets from the S-GWAS publication for lung cancer (n = 9,178 individuals), bladder cancer (n = 13,060) and age-related macular degeneration (AMD; n = 22,683) (Methods). Each dataset was divided into two subsets to simulate a joint study involving two cohorts that could not be combined. For all three datasets, we observed a substantial reduction in both runtime (Fig. 2a) and communication (Fig. 2b) for SF-GWAS compared to S-GWAS. Communication refers to the transfer of encrypted information between parties required during the interactive protocol, which is a key performance metric as its impact on overall runtime can vary depending on the network conditions in practice. SF-GWAS ran consistently an order of magnitude faster than S-GWAS (for example, 4.6 h versus 64.3 h for AMD data, representing a 14× reduction). Communication cost was three to four times lower for SF-GWAS (for example, 173.7 GB versus 666.6 GB for AMD data), largely because S-GWAS requires sharing the entire encrypted dataset between the parties, which SF-GWAS circumvents through our federated approach. The output of SF-GWAS closely matched a direct analysis of the pooled plaintext data (Extended Data Fig. 1). Furthermore, SF-GWAS provides stronger security properties than S-GWAS by avoiding the sharing of entire encrypted input datasets while adopting an encryption scheme that is resilient to quantum computer-based attacks (Supplementary Note).

We evaluated the scalability of SF-GWAS on two larger datasets: the eMERGE consortium (n = 31,293) and UK Biobank (UKB; n = 275,812), focusing on body mass index (BMI) as the analysis trait in both. These datasets are more than 100 and 2,000 times larger, respectively, than the largest AMD dataset analyzed in the S-GWAS publication, in part due to the number of variants analyzed (509,000 in AMD versus 38 and 93 million in eMERGE and UKB, respectively). We split both datasets across multiple centers (six for eMERGE and seven for UKB) according to the original data collection sites, considering each center's dataset as local and private (Methods). S-GWAS could not be evaluated on either dataset due to its infeasible runtime requirements, estimated through linear extrapolation to be several months for eMERGE and several years for UKB.

https://doi.org/10.1038/s41588-025-02109-1

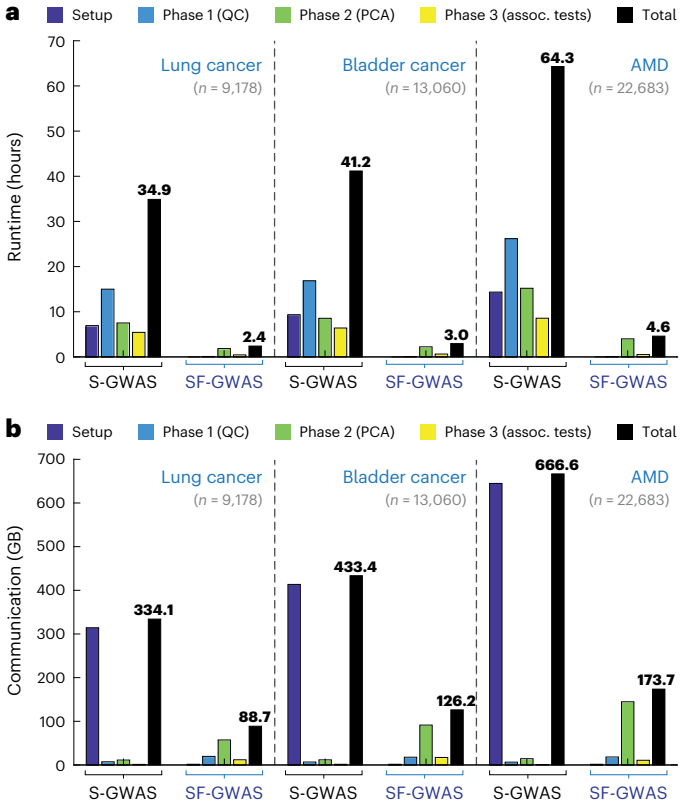


Fig. 2 | SF-GWAS is more computationally efficient than the prior art for secure collaborative GWAS. a,**b**, We compare the runtime (**a**) and communication cost (**b**) of SF-GWAS (PCA-based) with those of Secure GWAS[4] (S-GWAS). Unlike other existing cryptographic solutions for GWAS, these two methods implement the full pipeline including QC and PCA, both of which are standard steps of GWAS. We analyzed the three datasets evaluated in the S-GWAS publication for lung cancer, bladder cancer and AMD using linear regression-based association tests, following the previous work. Each dataset is evenly split into two parts and distributed between two machines to simulate a collaborative GWAS setting. In addition to the total costs (in black), we show the costs of individual steps, including the initial setup and the three phases (QC, PCA and association tests). The setup involves key generation and network connection for both methods, and additionally encrypted data transfer (secret sharing) for S-GWAS, which is not required by SF-GWAS due to its federated nature. For all datasets, SF-GWAS reduces the overall runtime by an order of magnitude and reduces the communication by a factor of around 3.5. We also note that SF-GWAS provides stronger security properties than S-GWAS by avoiding the sharing of entire encrypted input datasets while adopting an encryption scheme that is resilient to quantum computer-based attacks (Supplementary Note). Plots showing the accuracy of SF-GWAS results are provided in Extended Data Fig. 1. Unlike S-GWAS, SF-GWAS supports both linear and logistic regression-based tests, even for large-scale datasets. Extended Data Figs. 8 and 9 depict SF-GWAS's runtime, communication cost and accuracy for the latter.

Consistent with our previous results, the association statistics computed by SF-GWAS, for BMI and using the PCA-based pipeline, closely matched a plaintext analysis based on the PLINK software[12] on each of the pooled datasets (Fig. 3a,b and Extended Data Fig. 2). In addition, we observed that a meta-analysis of summary statistics from individual centers can result in notable discrepancies compared to the pooled analysis (Extended Data Fig. 3), especially on the eMERGE dataset, likely due to limited site-specific sample sizes and the heterogeneity of study populations[13–15]. Because the computational steps of SF-GWAS closely emulate a centralized analysis, SF-GWAS results are virtually equivalent to what the researchers would obtain if the datasets could be directly combined. This equivalence holds regardless of how the data are split (for example, even when the data distribution is heterogeneous).

The total runtime of SF-GWAS on eMERGE was 17.5 h, including 2.8 h for QC filtering, 8 h for PCA and 6.7 h for association tests. For UKB, the runtime was 5.3 days in total, including 4.5 h for QC, 44 h for PCA and 77.8 h for association tests. A concurrent work on multisite GWAS[16] reported a runtime of 5 h for 160,000 single-nucleotide polymorphisms (SNPs) and 16,000 samples, which extrapolates to more than 4 months for the UKB dataset, while providing weaker security guarantees than SF-GWAS and addressing only the PCA-based workflow. The runtime of PCA-based SF-GWAS scales linearly with dataset size and thus can be readily estimated in practice (Extended Data Fig. 4 and Supplementary Note). Furthermore, given the reduced reliance of SF-GWAS runtime on interactive steps as dataset size grows, we expect it to remain practical for international collaborations with higher network communication delays, such as those between the US and the UK (Extended Data Fig. 5).

Next, we evaluated our secure and federated algorithm for LMM-based association tests, based on the plaintext method of REGENIE[17], on a dataset of 409,548 individuals of European descent from UKB including related individuals (Methods). We split this dataset among six centers, as in the previous experiment. SF-GWAS produced association statistics that accurately matched those of REGENIE, where the latter was run directly on a pooled dataset without encryption (Fig. 3c and Extended Data Fig. 2). We additionally validated the accuracy of our method on the lung cancer dataset from S-GWAS (Extended Data Fig. 6). Owing to our optimizations, LMM-based SF-GWAS exhibits near-constant runtime scaling with the size of local datasets and achieves runtimes on the order of days, even for large datasets containing hundreds of thousands of individuals (Extended Data Fig. 7 and Supplementary Note). In our experiment, it achieved a runtime of 6 days for the UKB dataset.

GWAS of binary traits (for example, disease status) is typically performed using logistic regression, requiring iterative model fitting in the absence of an analytical solution. Given the high computational costs of nonlinear operations under encryption, existing works on secure GWAS have focused on linear regression or have been limited to supporting logistic models on small datasets[4,6]. SF-GWAS incorporates a practical, secure federated algorithm for score-based tests for logistic models using Newton's method, ensuring fast convergence to accurate parameter estimates (Methods and Extended Data Fig. 8). Reanalyzing the three S-GWAS datasets using our logistic workflow produces results consistent with PLINK and achieves practical runtimes of less than 5.3 hours for all datasets, comparable to those of the linear pipeline (Extended Data Figs. 8 and 9).

Collaborative analysis performed using SF-GWAS identified genetic variants that are concordant with prior GWAS results. Comparing with the published summary statistics from the Pan-UK Biobank project[18], we observed that 71 out of 73 significant variants ($P < 5 \times 10^{-8}$) identified by SF-GWAS on eMERGE coincided with previously reported associations for BMI, whereas independent analysis of each center's data resulted in at most two significant variants. Similarly for UKB, 1,778 out of 2,200 significant variants for LMM-based SF-GWAS, and 21,544 out of 24,357 for our PCA-based analysis (based on a larger set of imputed genotypes), were previously reported to be significant, indicating a large overlap despite differences in the analysis setting. In contrast, independently analyzing each center-specific dataset (PCA based) identified 2,600 significant variants across all centers, considerably fewer than SF-GWAS (24,357). Furthermore, when treating UKB as a validation cohort, meta-analyses of seven center-specific GWAS based on eMERGE data yielded fewer variants validated in UKB compared to SF-GWAS, illustrating the improved statistical power of the joint analysis enabled by SF-GWAS (Extended Data Fig. 10).

Finally, we demonstrate an application of SF-GWAS to analyzing datasets collected independently by different organizations. In addition to the AMD GWAS cohort from the International AMD Genomics Consortium (IAMDGC; $n = 21,692$, including 9,284 cases), we selected ancestry-matched (European) AMD cohorts in the eMERGE consortium

https://doi.org/10.1038/s41588-025-02109-1

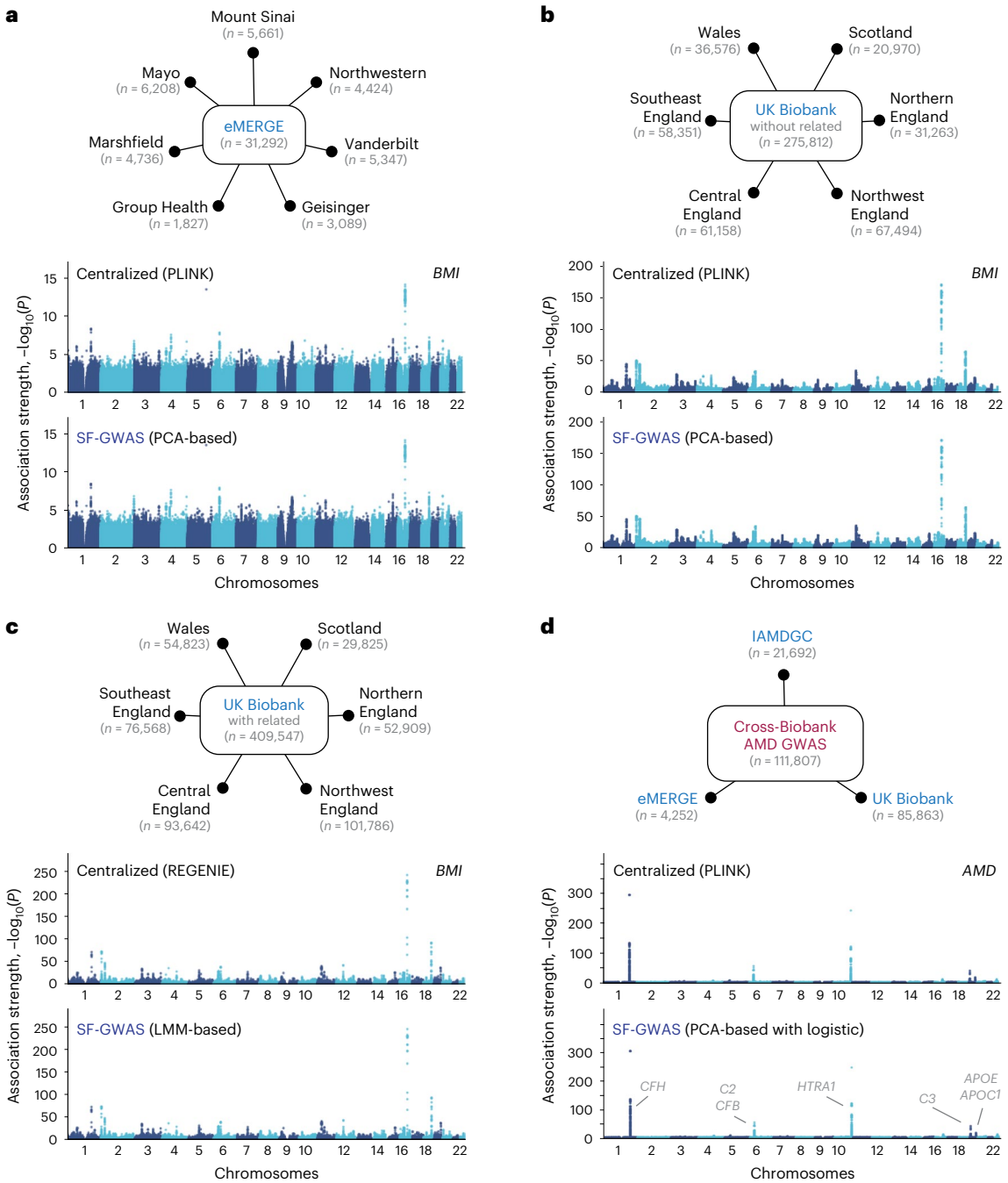


Fig. 3 | SF-GWAS accurately reproduces biobank-scale GWAS without data centralization. a–**d**, We evaluated SF-GWAS on eMERGE (**a**), UKB (**b**,**c**), and cross-biobank AMD GWAS (**d**) datasets to demonstrate its applicability to biobank-scale collaborative GWAS. Considering both the number of individuals and the number of variants (Methods), eMERGE is at least 100×, and both UKB and cross-biobank AMD are at least 2,000× larger than the largest dataset considered by the prior work S-GWAS[4]. The total sample count and the sizes of individual datasets used in the federated setting are shown (left); we split eMERGE data into seven groups according to the data providing organization, and for UKB, we geographically grouped the original sample collection sites into six groups. Cross-biobank AMD represents a setting where joint analysis is performed over a heterogeneous collection of independently collected GWAS datasets. Following standard practice, we excluded individuals with a close relative in the dataset for PCA-based GWAS (**b**), whereas the full cohort was considered for the LMM-based pipeline (**c**). For eMERGE and UKB (**a**–**c**), we assessed the genetic associations of BMI, accounting for covariates (age, sex and center) and principal components (5 for eMERGE, 10 for UKB), where the latter are globally computed over the entire dataset (securely performed in SF-GWAS). The *P* values were derived from a two-sided Wald test. For a cross-biobank GWAS of AMD (**d**), we performed a logistic regression-based tests with the same covariates. The *P* values were calculated using a two-sided score-based test for the logistic model. All *P* values shown, across all methods, were not adjusted for multiple testing. Manhattan plots visualizing the association strength of individual variants across chromosomes 1–22 are shown for a centralized, unencrypted analysis using PLINK or REGENIE software (for PCA-based and LMM-based workflows, respectively) and a secure and federated analysis using SF-GWAS. In all experiments, SF-GWAS accurately reproduces the corresponding centralized analysis without requiring the collaborating entities to share private data. Scatter plots comparing the association statistics are provided in Extended Data Figs. 2 and 9.

(n = 4,252, including 574 cases) and the UKB (n = 85,863, including 2,632 cases). This resulted in a GWAS dataset of 111,807 individuals in total, split among three parties representing the original data sources. Although the three cohorts were genotyped using different array platforms, a large number of shared genetic variants (22,191,946) could be chosen for joint analysis by imputing each dataset and finding an intersecting set of genomic positions and allele definitions. The total runtime of SF-GWAS (PCA-based, logistic) was 3 days, including 53 min for QC filtering, 8 h for PCA and 68 h for association tests. For comparison, we estimate a runtime of 42 days for an existing cryptographic solution for GWAS[6] (which does not support QC or PCA), assuming linear scaling with both the number of computing cores (inversely) and the number of variants. Consistent with previous experiments, SF-GWAS obtained accurate results (Fig. 3d and Extended Data Fig. 9), and the top five strongest associations identified by SF-GWAS correspond to loci and genes with well-established roles in AMD, including *CFH*, *C2*/*CFB*, *HTRA1*, *C3*, and *APOE*/*APOC1* in chromosomes 1, 6, 10, 19 and 19, respectively.

When using SF-GWAS in collaborative studies, data harmonization may be required to identify shared genetic variants across sites and unify allele or phenotype definitions. This can be facilitated by sharing non-private metadata, such as genomic positions, without exposing sensitive raw data. Notably, SF-GWAS facilitates cross-site variant QC by securely calculating global statistics, taking all parties' data into account. This can improve data quality, for instance, by rescuing rare variants that are too infrequent at individual sites. In addition, a recent method for securely detecting related individuals across sites could be incorporated to aid cohort selection[19]. Another important practical consideration is the ease of deployment with security assurances. To address this, we developed our open-source software to operate in any environment with minimal dependencies, and the sfkit web server[20] can be used to automate the deployment of our workflows using either cloud resources or private machines. Although server networking may require organizational approval, SF-GWAS's federated design and encrypted sharing of information minimize risks and support regulatory compliance[21].

In summary, our work demonstrates a secure and federated approach to multisite GWAS with rigorous privacy guarantees and scalable performance. Further extending our methods to other useful analysis tasks, including logistic mixed-effects models and statistical corrections for imbalanced or biased datasets[22–24], as well as ensuring security even against malicious actors who may deviate from the prescribed protocols, are meaningful directions for future work. We expect our algorithmic techniques and open-source software to accelerate the development of secure genomic data analysis methods. In a global environment where there are increasing concerns and legal restrictions over sharing of sensitive data about individuals, our work lays the foundation for future biomedical advances by facilitating cross-institutional collaboration.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-025-02109-1.

References

1. Graham, S. E. et al. The power of genetic diversity in genome-wide association studies of lipids. *Nature* **600**, 675–679 (2021).
2. Trubetskoy, V. et al. Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**, 502–508 (2022).
3. Wan, Z. et al. Sociotechnical safeguards for genomic data privacy. *Nat. Rev. Genet.* **23**, 429–445 (2022).
4. Cho, H., Wu, D. J. & Berger, B. Secure genome-wide association analysis using multiparty computation. *Nat. Biotechnol.* **36**, 547–551 (2018).
5. Froelicher, D. et al. Truly privacy-preserving federated analytics for precision medicine with multiparty homomorphic encryption. *Nat. Commun.* **12**, 5910 (2021).
6. Blatt, M., Gusev, A., Polyakov, Y. & Goldwasser, S. Secure large-scale genome-wide association studies using homomorphic encryption. *Proc. Natl Acad. Sci. USA* **117**, 11608–11613 (2020).
7. Gürsoy, G. et al. Functional genomics data: privacy risk assessment and technological mitigation. *Nat. Rev. Genet.* **23**, 245–258 (2022).
8. Berger, B. & Cho, H. Emerging technologies towards enhancing privacy in genomic data sharing. *Genome Biol.* **20**, 128 (2019).
9. Arellano, A. M., Dai, W., Wang, S., Jiang, X. & Ohno-Machado, L. Privacy policy and technology in biomedical data science. *Annu. Rev. Biomed. Data Sci.* **1**, 115–129 (2018).
10. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
11. Astle, W. & Balding, D. J. Population structure and cryptic relatedness in genetic association studies. *Stat. Sci.* **24**, 451–471 (2009).
12. PLINK software (accessed 31 January 2024); https://www.cog-genomics.org/plink/
13. Kanai, M. et al. Meta-analysis fine-mapping is often miscalibrated at single-variant resolution. *Cell Genom.* **2**, 100210 (2022).
14. Boedhoe, P. S. et al. An empirical comparison of meta-and mega-analysis with data from the enigma obsessive-compulsive disorder working group. *Front. Neuroinform.* **12**, 102 (2019).
15. Nasirigerdeh, R. et al. sPLINK: a hybrid federated tool as a robust alternative to meta-analysis in genome-wide association studies. *Genome Biol.* **23**, 32 (2022).
16. Yang, M. et al. TrustGWAS: A full-process workflow for encrypted GWAS using multi-key homomorphic encryption and pseudorandom number perturbation. *Cell Syst.* **13**, 752–767 (2022).
17. Mbatchou, J. et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).
18. Pan-UK Biobank project (Pan-UK Biobank, 2020); https://pan.ukbb.broadinstitute.org/
19. Hong, M. M. et al. Secure discovery of genetic relatives across large-scale and distributed genomic data sets. *Genome Res.* **34**, 1312–1323 (2024).
20. Mendelsohn, S. et al. sfkit: a web-based toolkit for secure and federated genomic analysis. *Nucleic Acids Res.* **51**, W535–W541 (2023).
21. Scheibner, J. et al. Revolutionizing medical data sharing using advanced privacy-enhancing technologies: technical, legal, and ethical synthesis. *J. Med. Internet Res.* **23**, e25120 (2021).
22. Wang, X. Firth logistic regression for rare variant association tests. *Front. Genet.* **5**, 187 (2014).
23. Dey, R., Schmidt, E. M., Abecasis, G. R. & Lee, S. A fast and accurate algorithm to test for binary phenotypes and its application to phewas. *Am. J. Human Genet.* **101**, 37–49 (2017).
24. Beesley, L. J. & Mukherjee, B. Statistical inference for association studies using electronic health records: handling both selection bias and outcome misclassification. *Biometrics* **78**, 214–226 (2022).

Methods

Ethics and inclusion statement

The use of all controlled-access datasets in this study were approved by the respective data access committees through the NIH Database of Genotypes and Phenotypes (dbGaP) and the UKB Access Management System (project ID: 46341 and 41910).

Review of secure MPC

MPC techniques enable multiple entities to securely and interactively perform computation on private inputs (Supplementary Note). Standard MPC frameworks[25] leverage (additive) secret sharing, where each private value is split into random (encrypted) shares, which are in turn distributed to different computing parties. Although the shares collectively encode the private value, any subset of shares provably does not leak any private information. Computing parties then collaborate and use the secret shares to evaluate a function on the private input without revealing information about the private input to any entity involved. For example, the secure addition of two secretly shared numbers x and y can be executed by having each party add their individual shares for x and y. The new shares constitute a sharing of $x+y$, which is the desired computation result. More sophisticated functions (for example, multiplication, division, square root and sign) can be similarly defined over the secret shares but require the two parties to interact by exchanging a sequence of numbers, which also do not reveal private information. These secure routines can be composed to perform arbitrary operations on private input data held by multiple entities. However, the communication cost of MPC can introduce a bottleneck in applications involving complex tasks. In addition, secret sharing requires that the entire input data be shared with all computing parties.

Review of HE

HE is a form of encryption that allows for direct computations over encrypted data, without having to decrypt them. Initially developed for limited types/rounds of operations, HE is now widely applicable to many analysis tasks due to the recent introduction of fully HE schemes, which include a *bootstrapping* routine to allow an arbitrary number of operations to be performed, and the development of efficient techniques for common scientific operations. For instance, the CKKS scheme[26] encodes a vector of continuous values in a single ciphertext and is well-suited for calculations where a small amount of noise can be tolerated. Like other HE schemes, CKKS performs both additions and multiplications simultaneously on the encrypted values within a ciphertext (single instruction, multiple data property), which improves the overall scalability of the scheme. While HE uniquely enables a single party to perform computation on the encrypted data without interaction, the computational cost and flexibility of HE remain more limited than MPC for general tasks. Also, for multisite collaboration, one needs to transfer all of the encrypted data to a single machine for joint analysis, which can be challenging for large datasets.

Combining HE and MPC to enable practical, secure federated computation

To address the performance limitations of HE and MPC, we take a federated approach to secure computation leveraging both HE and MPC, where the input datasets are kept locally at the respective source sites and only small intermediate data are securely exchanged among the parties using cryptographic techniques to carry out a global computation. For the HE component, we build upon a multiparty HE (MHE) scheme (related to threshold HE) based on CKKS, which extends the CKKS scheme to the setting with multiple data providers by secret sharing the decryption key and constructing a shared encryption key (Supplementary Note). Under our scheme, any party can encrypt the data and perform HE computations locally, but decryption can be performed only if all parties cooperate. Our approach allows each party to perform local computations involving the unencrypted input dataset and a small amount of encrypted data, whereas certain global computations are performed by sharing intermediate results among the parties, encrypted under the shared encryption key. At the end of the protocol, all parties collaborate to decrypt the final results. By keeping each input dataset local, we minimize the communication and enable local plaintext computation, which are substantially more efficient than corresponding ciphertext computation. We also leverage an efficient interactive protocol for bootstrapping in MHE[27] to reduce the overall computational burden of HE.

Improving upon existing works on MHE[5,28–30], we switch between the MHE and secret sharing representations of intermediate data, which enables efficient MPC routines to be used in conjunction with MHE operations to carry out the global computations (Supplementary Note). We convert between the two schemes to perform each operation under the most efficient scheme throughout the GWAS pipeline. For example, we perform large-scale matrix and vector operations using MHE to exploit its SIMD (single instruction, multiple data) property, but evaluate non-polynomial functions (division, square root and comparison) with compact bit-wise MPC protocols, which are more efficient and numerically stable than the MHE counterparts. Any operation involving the local unencrypted data is performed using MHE to avoid secret sharing of the large input datasets.

Our algorithmic design strategies for enabling secure population structure correction

Our federated framework for secure computation allows us to develop efficient and provably secure algorithms for collaborative GWAS. Our work introduces practical methods for two standard approaches to account for population structure, namely PCA and LMMs. We adopt the following algorithmic design strategies to obtain accurate and efficient performance. First, we closely adhere to the computational pipeline of the desired centralized algorithm to obtain accurate results while securely operating over private datasets held by multiple parties. This is in contrast to other collaborative approaches that simplify or approximate the analysis to address the lack of access to a pooled dataset (for example, meta-analysis). Next, we restructure the computation while maintaining its equivalence to the original algorithm to both maximize the use of low-cost operations and minimize communication by leveraging local plaintext data. We switch between MPC and MHE routines throughout the protocol to improve the efficiency and numerical robustness of our routines. We also optimize the vectorized encoding of data in encrypted representations for efficient composition of linear algebra operations. We detail our algorithmic strategies and optimization techniques in Supplementary Note.

In addition to enabling a substantial performance improvement for PCA compared to prior work[4], our techniques facilitate the design of a practical protocol for secure and federated LMMs (Supplementary Note). LMMs typically involve operations on the genetic relatedness matrix, which scale with the number of individuals in the dataset and impose a notable computational burden for large cohorts, even in centralized analysis settings. Our approach builds upon REGENIE[17], a recently developed algorithm for LMM association tests based on a stacked ridge regression approach, which directly models the ancestry-related confounding effect as the output of a genome-wide regression model, thus circumventing the use of a genetic relatedness matrix. This approach brings scalability improvements while providing accuracy comparable to other LMM methods such as BOLT-LMM[31], fastGWA[32] and SAIGE[33]. Unfortunately, REGENIE cannot be directly applied in a secure federated setting. Although ridge regression can be efficiently performed in plaintext on a pooled dataset, implementing standard algorithms (for example, those based on the closed-form solution) with secure computation techniques results in impractical runtime requirements due to the complex matrix operations, such as inversion, that need to performed on large encrypted matrices.

We address this challenge by comprehensively redesigning REGENIE's stacked regression procedure with our secure federated algorithms for ridge regression based on conjugate gradient descent and alternating direction method of multipliers (ADMM). Although the inclusion of private covariate features makes a direct application of the latter approach[34] impractical in our setting, we reformulated the ADMM algorithm (referred to as ADMM-Woodbury) by leveraging a matrix identity to enable plaintext precomputation of computationally expensive operations (for example, large-matrix inversion) at each site. This reformulation maximizes efficient local computation and, consequently, achieves near-constant runtime scaling with respect to cohort size (Extended Data Fig. 7).

Our scalable techniques for secure association tests based on linear and logistic models

Performing association tests at genome-wide scale involves intricate linear algebra operations (for example, projection, inversion, and factorization) applied to large-scale matrices and vectors. Following the strategies in the previous section, we designed efficient algorithms for secure federated association tests that can analyze millions of genetic variants within practical runtimes. For example, we developed optimized routines for matrix multiplications between the local genotype matrix and an encrypted low-dimensional matrix of covariates and phenotypes by identifying and precomputing intermediate terms that involve the unencrypted local dataset, which can be more efficiently computed than operations involving only encrypted data. This approach also helps to distribute the workload among the parties according to the size of the local datasets. For complex, iterative operations such as matrix factorization and inversion, we securely convert between MHE and MPC frameworks to take advantage of efficient MPC routines for costly operations such as comparison and division. Furthermore, after the initial precomputation, we process genetic variants in non-overlapping genomic blocks in parallel to accelerate the computation.

Prior to our work, logistic regression-based association tests were considered more challenging and could be performed using cryptographic techniques only on small datasets. This limitation has mainly been due to the costly iterative optimization procedure required to estimate the coefficients of logistic regression models (for example, genetic effect sizes). To address this challenge, we developed an efficient secure federated algorithm for score-based tests in logistic models (Supplementary Note). Our algorithm securely obtains statistics that are asymptotically equivalent to alternative tests (that is, Wald and likelihood-ratio) and can be applied to large biobank datasets. Unlike the Wald or likelihood-ratio tests, both of which require a separate logistic regression model to be fitted for every variant being tested, score-based tests fit a single null model including only the covariates, subsequently allowing all variants to be tested using a closed-form expression without refitting the model. To obtain the initial null model efficiently and accurately, we devised a federated algorithm for Newton's method harnessing both MHE and MPC for efficient matrix multiplications and Hessian matrix inversion, respectively. This approach enables faster convergence to precise parameter estimates compared to standard first-order gradient-based methods (Extended Data Fig. 8). Once the null model is obtained, the remaining computation of association statistics mostly consists of linear algebraic operations, which can be performed efficiently using our secure federated approach. As our results show, these techniques enable SF-GWAS to support accurate logistic regression-based association tests with practical runtimes comparable to those of the linear regression setting.

SF-GWAS pipelines

Our SF-GWAS algorithm implements the full GWAS pipeline, including QC, correction for population structure (PCA and LMM) and association tests. Collaborating parties first agree on the phenotype, covariates and a list of genetic variants to analyze, as well as study parameters, such as filtering thresholds and other algorithmic parameters. They also agree on the security parameters and generate the required cryptographic keys for HE and MPC, for example, encryption keys and shared pseudorandom number generators (Supplementary Note). They then proceed with the interactive protocol to securely perform the desired computation. For QC, the parties independently filter their subset of samples based on public thresholds (for example, for heterozygosity and missing genotype rate), then utilize MPC routines to jointly and securely filter the variants based on global statistics (for example, minor allele frequencies and Hardy-Weinberg equilibrium). The variant filtering output is shared with all parties so that the rest of the protocol can proceed with the reduced variant set.

For the PCA-based workflow, we implemented a secure federated algorithm for randomized PCA[4,35] to compute the top principal components (PCs) over the entire dataset without constructing the pooled matrix (Supplementary Note). The jointly computed PCs are kept encrypted for use in the following steps. For linear regression-based association tests, we use both HE and MPC operations to compute the covariate-corrected association statistics based on a linear model. This step includes (i) a secure federated QR factorization for computing the joint orthogonal basis of PC and the observed covariates, (ii) secure matrix multiplication based on HE to project the genotypes onto the covariate subspace for correction and to compute genotype-phenotype covariances and (iii) MPC routines to inversely scale the statistics by the standard deviations of the genotypes and the phenotype for normalization.

For binary traits, we use our secure federated algorithm for score-based tests using logistic regression models (Supplementary Note). The first step is to fit a null logistic regression model across the parties considering only the covariates, for which we introduce a federated Newton's method based on both HE and MPC. Then, we compute the components of the score-based test statistic with respect to each tested variant, which include (i) the score (that is, the derivative of the log-likelihood function with respect to the coefficient of the genetic effect) and (ii) the Fisher information (related to the standard error of the coefficient). Both terms and the resulting statistics are computed securely in a federated manner using a series of linear algebra operations in HE and nonlinear operations in MPC. Finally, the association statistics are collectively decrypted and shared among the parties as the final output of the analysis.

For the LMM-based workflow, we implemented a secure federated version of REGENIE[17] that is based on stacked ridge regression models (Supplementary Note). After the QC step, the genetic variants are first grouped into fixed-size blocks. For each block, a ridge regression model is jointly trained across the parties using our ADMM-Woodbury algorithm to obtain encrypted phenotype predictions, which leverage only the variants within the block while accounting for linkage disequilibrium. Subsequently, the block-wise local predictions are provided as input features to a genome-wide ridge regression model for phenotype prediction, jointly trained across the parties using our conjugate gradient descent algorithm. We adopt a cross-validation scheme to determine an appropriate choice of variance parameter, representing the genomic heritability of the target phenotype. We selected K-fold cross-validation, as it is typically more efficient than other methods such as leave-one-out cross-validation and achieves almost identical accuracy in this framework[17]. Association tests are performed by calculating the correlation between each target variant and the phenotype residuals, excluding the contribution from the genome-wide regression model, which is estimated without the chromosome containing the tested variant. Similar to the PCA-based workflow, we compute the association statistics efficiently using optimized secure matrix multiplication routines combined with MPC routines for data normalization.

Further details on the SF-GWAS pipelines, including their security, runtime and complexity analyses, are provided in Supplementary Note and Supplementary Tables 2–5.

Related work

Several studies have proposed methods for securely performing collaborative GWAS over private datasets using secure computation frameworks such as HE or MPC[4–6]. However, these approaches are often limited by runtime constraints or the inability to perform GWAS analyses while accounting for population structure or sample relatedness. MPC frameworks based on secret sharing[4] are particularly communication-intensive and require encrypting and distributing the entire dataset among all participating parties, resulting in substantial overhead. Although HE enables noninteractive computation on encrypted data, joint analysis using HE[6] still requires collaborating entities to centralize the encrypted data for analysis by a single party. This approach leads to impractical computational costs for complex analysis tasks on large genomic datasets. Prior work applying MHE to GWAS[5] addresses some limitations of centralized HE but is restricted to association testing, excluding other critical GWAS components such as QC filtering and correction for population structure (PCA or LMMs). Furthermore, this earlier work relied solely on HE computations, without integrating MPC protocols to improve accuracy or enhance efficiency for nonpolynomial operations. Secure hardware-based approaches to GWAS, such as those based on Intel SGX[36], have also been proposed. However, these methods are also limited to association testing without incorporating PCA or LMMs and lack the formal privacy guarantees provided by HE or MPC, making them susceptible to known security risks[37–39].

Another approach to collaborative GWAS is based on federated learning techniques, which iteratively aggregate intermediate results among parties to perform global computations[15,40]. These solutions are generally more accurate than meta-analysis, as they can more closely emulate a pooled analysis[15]. They also achieve efficient performance because plaintext (unencrypted) data are locally held and directly analyzed by each party. However, these methods require intermediate results to be shared among the parties (or with a trusted third party) in plaintext, often involving many rounds of interaction, which increases the risk of privacy leakage[41,42]. Although differential privacy techniques can mitigate such leakage, existing methods are not practical for GWAS, as releasing statistics for a large number of variants would require adding substantial noise to preserve privacy[43]. In SF-GWAS, parties keep their local data in plaintext while exchanging only encrypted intermediate results throughout the study. This approach efficiently emulates a pooled analysis in a distributed fashion while ensuring formal privacy guarantees.

Benchmark datasets

We obtained the three datasets used in the original S-GWAS publication[4] for comparison. These include a lung cancer dataset (n = 9,178, including 5,088 cases; 612,794 SNPs; females across all age groups), a bladder cancer dataset (n = 13,060, including 9,684 cases; 566,620 SNPs; both sexes across all age groups) and an AMD dataset (n = 22,683, including 6,211 cases; 508,740 SNPs; both sexes across all age groups). We followed the steps in the prior work to prepare the data and then evenly and uniformly split each dataset between two parties to emulate a multisite study.

For the eMERGE data, we obtained a cohort of 31,292 individuals split across seven study groups: Geisinger Health System (n = 3,089), Group Health (University of Washington; n = 1,827), Marshfield Clinic (Pennsylvania State University; n = 4,736), Mayo Clinic (n = 6,208), Icahn School of Medicine at Mount Sinai (n = 5,661), Northwestern University (n = 4,424) and Vanderbilt University (n = 5,347). The individuals range in age from 3 to 75 years and include both sexes. We used a total of 38,040,168 imputed biallelic SNPs for the analysis and chose BMI as the target phenotype. Four covariates were included in the analysis: membership to each study group, age (at time of assessment), sex and age^2.

For the UKB data, we obtained a cohort of 409,547 individuals of European descent, including both sexes, aged 40 to 69. For use with the PCA-based GWAS pipeline, we also constructed a subset of 275,812 unrelated individuals (King[44] relatedness coefficient less than 0.062). This cohort represented 22 different health centers across the United Kingdom. To simulate a federated study, we grouped the health centers into six study groups based on geographic location: Scotland (n = 29,825 in total; 20,970 unrelated), Northern England (n = 52,909; 31,263), Northwest England (n = 101,786; 67,494), Central England (n = 93,642; 61,158), Southeast England (n = 76,568; 58,351) and Wales (n = 54,823; 36,576). We provide the list of centers in each group in Supplementary Table 1. We used a total of 92,248,310 imputed biallelic SNPs for the PCA-based pipeline, and a subset of 581,927 non-imputed genotyped SNPs for the stacked regression models in the LMM-based pipeline, following the recommendation in the REGENIE software documentation[45]. We analyzed BMI as the target phenotype. Six covariates were included in the analysis: membership to each study group, age (at time of assessment), sex, age × sex, age^2, and age^2 × sex.

For the cross-biobank AMD GWAS analysis, we used data from three independent sources: eMERGE (n = 4,252, including 574 cases), UKB (n = 85,863, including 2,632 cases), and International AMD Genomics Consortium (IAMDGC; n = 21,692, including 9,284 cases) for a total of 111,807 individuals (12,490 cases). We used the imputed genotype data provided by eMERGE and UKB. For IAMDGC, we used Michigan Imputation Server (version 1) to impute the samples using the Haplotype Reference Consortium reference panel. The total number of imputed biallelic variants shared among the three datasets was 22,191,946. For UKB, AMD cases were ascertained based on the ICD-10 code H35.3, and the control samples were randomly sampled from the remaining cohort. For eMERGE and IAMDGC, we retained the AMD case and control labels provided in the original datasets. To demonstrate the scalability of our approach, we considered individuals of European ancestry in each dataset, determined based on self-reported ethnicity for UKB and GrafPop (version 1.0)-inferred ancestry for eMERGE and IAMDGC (P_e > 90%). Four covariates were included in the analysis: membership to each study group (two independent features), age (at time of assessment) and age^2. Other covariates of potential interest, such as sex, were not available in all datasets and thus were excluded from the analysis.

GWAS details

For the lung cancer, bladder cancer, and AMD datasets, we used the same QC filters as applied in the prior analysis of these datasets using S-GWAS[4]. For the eMERGE data, we used the following QC parameters: genotype missing rate per SNP < 0.1, minor allele frequency > 0.05, and Hardy-Weinberg equilibrium chi-squared test statistic < 23.928 (P > 10^{-6}). We used the same set of filters for the UKB and cross-biobank AMD data, except for minor allele frequency > 0.001 to account for the larger size of these datasets.

For PCA-based GWAS, we adopt the standard approach of using a reduced set of SNPs with low levels of linkage disequilibrium for the PCA step. SF-GWAS achieves this by imposing a minimum pairwise distance threshold of 100 kb after QC filtering, which we found to obtain similar results to alternatives based on a direct linkage disequilibrium calculation. To establish parity between the plaintext, centralized analysis and our SF-GWAS approach, we use the same set of SNPs for PCA in both analyses for our main results. Alternatively, SNP selection for PCA can be performed separately and the agreed-upon list of SNPs may be provided as input to SF-GWAS. For lung cancer, bladder cancer, AMD, eMERGE, and cross-biobank AMD datasets, we kept the top five PCs as covariates for the subsequent analysis. For UKB, we kept the top ten PCs. We assess the association between each SNP and phenotype of interest based on both linear and logistic regression models including covariates.

For linear association tests, SF-GWAS first constructs an orthogonal basis (Q) for the subspace defined by the covariates (for example, top principal components, age, sex, study group memberships), then computes the Pearson correlation coefficient (r) between the genotype and phenotype vectors where the covariate effects have been projected out using Q. The coefficient r for each SNP is revealed to the collaborating entities. The corresponding χ^2 statistic with one degree of freedom is obtained as $r^2(n-c)/(1-r^2)$, based on which a P value can be calculated. n and c denote the total number of individuals and the number of covariates, respectively. Note that this mapping does not reveal any additional information other than r. For logistic association tests, SF-GWAS computes the score-based test t-statistic for each SNP i as $t = (\tilde{\mathbf{g}}_i^T(\mathbf{y} - \hat{\mathbf{p}}))/\sqrt{\tilde{\mathbf{g}}_i^T \mathbf{W} \tilde{\mathbf{g}}_i}$, where $\tilde{\mathbf{g}}_i$ is the genotype residual vector across individuals after adjusting for covariates, $\mathbf{y}$ the phenotype vector, $\hat{\mathbf{p}}$ the vector of estimated mean of the trait based on the null model, and $\mathbf{W} = \mathbf{diag}\{\hat{p}_j(1-\hat{p}_j)\}$ with $\hat{p}_j$ the jth element of $\hat{\mathbf{p}}$ corresponding to the jth individual, following the approach of REGENIE[17]. The corresponding P value is estimated via a normal approximation with $t^2 \sim \chi_1^2$.

For LMM-based GWAS, we follow REGENIE's approach to first apply ridge regression to obtain the best predictive model of the covariate-corrected phenotype within a given genomic block, accounting for genotype correlations, and then perform a second regression to obtain genome-wide phenotype predictions based on the block-wise predictors. The size of each genomic block is set to 8,192 in order to maximally leverage the vectorized encryption scheme based on our cryptographic parameters. Following REGENIE, we use a five-fold cross-validation to select the best variance parameter to construct the final predictors. For the association tests, we adopt the standard leave-one-chromosome-out scheme, leaving out the chromosome including the tested variant to correct for the background genetic effect on the phenotype without interfering with the signal being tested. We then obtain the χ^2 statistic with one degree of freedom, analogous to the PCA-based pipeline, using the residuals of the leave-one-chromosome-out genome-wide predictors for each variant.

Evaluation approaches

We evaluated SF-GWAS by simulating each party on a separate virtual machine (VM) with 16 virtual CPUs (vCPUs) and 128 GB of memory (e2-highmem-16) on the Google Cloud Platform. For the main results, we adopt the most efficient local area network setting by creating the VMs within the same zone in the Google Cloud Platform; we illustrate the reasonable additional cost of wide-area network setting in Extended Data Fig. 5. For the UKB and cross-biobank AMD analyses, we used the same or larger VM types to utilize more vCPUs in parallel given the large size of the dataset. Specifically, we used n2-highmem-64 (64 vCPUs with 512 GB of RAM) and n2-highmem-128 (128 vCPUs with 864 GB of RAM) depending on the size of the local dataset.

We measured the runtime and communication cost of SF-GWAS (the latter measured by the number of bytes sent among the parties) on all datasets. We compared these metrics against the prior work implementing the analogous PCA-based GWAS pipeline, S-GWAS[4]. We also evaluated SF-GWAS's scaling with respect to the number of samples, SNPs, covariates and computing parties (Extended Data Fig. 4). For the scaling experiments, we replicated the lung cancer dataset to produce a dataset of desired dimensions and modified GWAS parameters as needed to ensure that the amount of data at each step grew proportionally with the input dimensions (for example, ensuring that the number of samples passing QC doubles when the original number of samples doubles).

To evaluate the accuracy of SF-GWAS, we compared its association statistics to those obtained from a plaintext, centralized analysis where the individual datasets are combined to form a single consolidated dataset for analysis. For the three S-GWAS datasets, we used a plaintext Python (version 3.8) implementation of the same procedure as SF-GWAS (using a standard PCA implementation in the scikit-learn package[46], version 1.3.0); and for eMERGE, UKB, and the cross-biobank AMD analysis, we used the PLINK software (version 2.0; https://www.cog-genomics.org/plink/2.0/) implementing the same pipeline for PCA-based GWAS. For LMM-based GWAS, we used the REGENIE software (version 2.2.4; https://rgcgithub.github.io/regenie/) on the pooled dataset with the same parameters to obtain the ground truth association results. For eMERGE and UKB, we additionally evaluated the accuracy of meta-analysis approaches by performing a separate GWAS for each study group with the same study parameters as the global analysis and then combining the association statistics among different parties using the meta-analysis methods implemented in PLINK.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The three datasets used for comparison with the prior work on Secure GWAS[4] are available via NIH dbGaP with accession numbers phs000716.v1.p1 (lung cancer), phs000346.v2.p2 (bladder cancer) and phs001039.v1.p1 (AMD; International AMD Genomics Consortium [IAMDGC]). The eMERGE consortium data are also available via dbGaP (phs000888.v1.p1). Data access applications for the UKB data can be submitted at https://www.ukbiobank.ac.uk/.

Code availability

Our open-source software for SF-GWAS, which includes data preprocessing, analysis and plotting scripts to reproduce the main GWAS results of this publication, is available on GitHub (https://github.com/hhcho/sfgwas) and Zenodo at https://doi.org/10.5281/zenodo.14726447 (ref. 47). In addition, automated workflows for SF-GWAS are available on the sfkit web server[20] (https://sfkit.org), which allows a group of users to perform a secure joint analysis of their private datasets using either cloud computing resources or private machines.

References

25. Keller, M. MP-SPDZ: a versatile framework for multi-party computation. In *Proc. ACM SIGSAC Conference on Computer and Communications Security (CCS)* 1575–1590 (ACM, 2020).
26. Cheon, J. H., Kim, A., Kim, M. & Song, Y. Homomorphic encryption for arithmetic of approximate numbers. In *Proc. International Conference on the Theory and Application of Cryptology and Information Security (ASIACRYPT)* 409–437 (ASIACRYPT, 2017).
27. Mouchet, C., Troncoso-Pastoriza, J. R., Bossuat, J.-P. & Hubaux, J. P. Multiparty homomorphic encryption from ring-learning-with-errors. In *Proc. Privacy Enhancing Technologies Symposium* 291–311 (IACR, 2021).
28. Froelicher, D. et al. Scalable privacy-preserving distributed learning. In *Proc. Privacy Enhancing Technologies Symposium* 323–347 (Sciendo, 2021).
29. Sav, S., Bossuat, J.-P., Troncoso-Pastoriza, J. R., Claassen, M. & Hubaux, J.-P. Privacy-preserving federated neural network learning for disease-associated cell classification. *Patterns* **3**, 100487 (2022).
30. Sav, S. et al. POSEIDON: privacy-preserving federated neural network learning. In *Proc. Network and Distributed Systems Security (NDSS) Symposium* (NDSS, 2021).
31. Loh, P.-R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
32. Jiang, L. et al. A resource-efficient tool for mixed model association analysis of large-scale data. *Nat. Genet.* **51**, 1749–1755 (2019).

33. Zhou, W. et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).
34. Zheng, W., Popa, R. A., Gonzalez, J. E. & Stoica, I. Helen: maliciously secure coopetitive learning for linear models. In *Proc. IEEE Symposium on Security and Privacy (S&P)* 724–738 (IEEE, 2019).
35. Halko, N., Martinsson, P.-G. & Tropp, J. A. Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.* **53**, 217–288 (2011).
36. Kockan, C. et al. Sketching algorithms for genomic data analysis and querying in a secure enclave. *Nat. Methods* **17**, 295–301 (2020).
37. Lipp, M. et al. PLATYPUS: software-based power side-channel attacks on x86. In *Proc. IEEE Symposium on Security and Privacy (S&P)* 355–371 (IEEE, 2021).
38. Van Bulck, J., Weichbrodt, N., Kapitza, R., Piessens, F. & Strackx, R. Telling your secrets without page faults: Stealthy page table-based attacks on enclaved execution. In *Proc. USENIX Security Symposium* 1041–1056 (USENIX, 2017).
39. Wang, W. et al. Leaky cauldron on the dark land: understanding memory side-channel hazards in SGX. In *Proc. ACM SIGSAC Conference on Computer and Communications Security (CCS)* 2421–2434 (ACM, 2017).
40. Zhu, R. et al. Privacy-preserving construction of generalized linear mixed model for biomedical computation. *Bioinformatics* **36**, i128–i135 (2020).
41. Melis, L., Song, C., De Cristofaro, E. & Shmatikov, V. Exploiting unintended feature leakage in collaborative learning. In *Proc. IEEE Symposium on Security and Privacy (S&P)* 691–706 (IEEE, 2019).
42. Zhu, L., Liu, Z. & Han, S. Deep leakage from gradients. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)* 32 (NIPS, 2019).
43. Simmons, S., Sahinalp, C. & Berger, B. Enabling privacy-preserving GWASs in heterogeneous human populations. *Cell Syst.* **3**, 54–61 (2016).
44. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
45. REGENIE: program for whole genome regression modelling of large genome-wide association studies (accessed November 2023); https://rgcgithub.github.io/regenie/
46. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
47. Cho, H., Froelicher, D., Chen, J. & Edupalli, M. SF-GWAS software and supporting scripts (v0.1.2). *Zenodo* https://doi.org/10.5281/zenodo.14726447 (2025).

Acknowledgements

We thank S. Mendelsohn and D. Loginov for integrating SF-GWAS into the sfkit web server. We thank members of the Center for Admixture Science and Technology for input regarding the deployment of our methods. This work was supported by National Institutes of Health (NIH) R01 HG010959 (to B.B.) and DP5 OD029574 and RM1 HG011558 (to H.C.). Part of this work was completed while H.C. was at the Broad Institute, supported by the Schmidt Fellows Program. In addition, this work was partially developed within the framework of the Data Protection in Personalized Health (DPPH) project (https://dpph-ch.github.io/), supported by the Swiss Personalized Health and Related Technologies (PHRT) Strategic Focus Area. This research has been conducted using data from the UKB, a major biomedical database (project ID: 46341 and 41910). We thank the study participants of datasets analyzed in this work.

Author contributions

H.C., D.F., A.P., J.R.T.-P., J.-P.H. and B.B. conceived the project. H.C., D.F., J.C., J.-P.H. and B.B. developed the methods. D.F., J.C., M.E. and H.C. implemented the software and performed experiments. H.C., D.F. and J.C. wrote the initial draft of the paper. All authors reviewed the results and edited the paper. H.C., J.-P.H. and B.B. guided the work.

Competing interests

J.R.T.-P. and J.-P.H. are co-founders of the start-up Tune Insight (https://tuneinsight.com). The other authors declare no competing interests.

Additional information

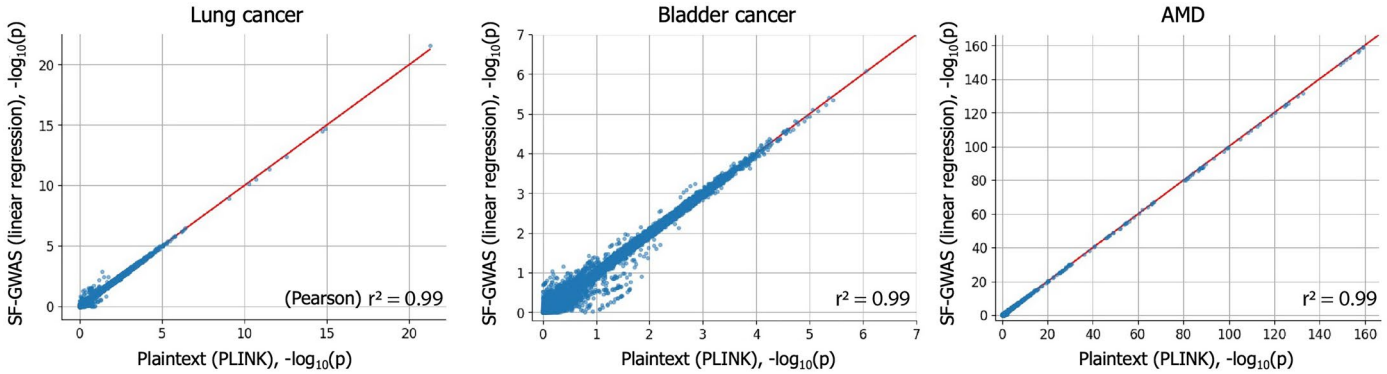
Extended data is available for this paper at https://doi.org/10.1038/s41588-025-02109-1.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-025-02109-1.

Correspondence and requests for materials should be addressed to Hyunghoon Cho, Jean-Pierre Hubaux or Bonnie Berger.

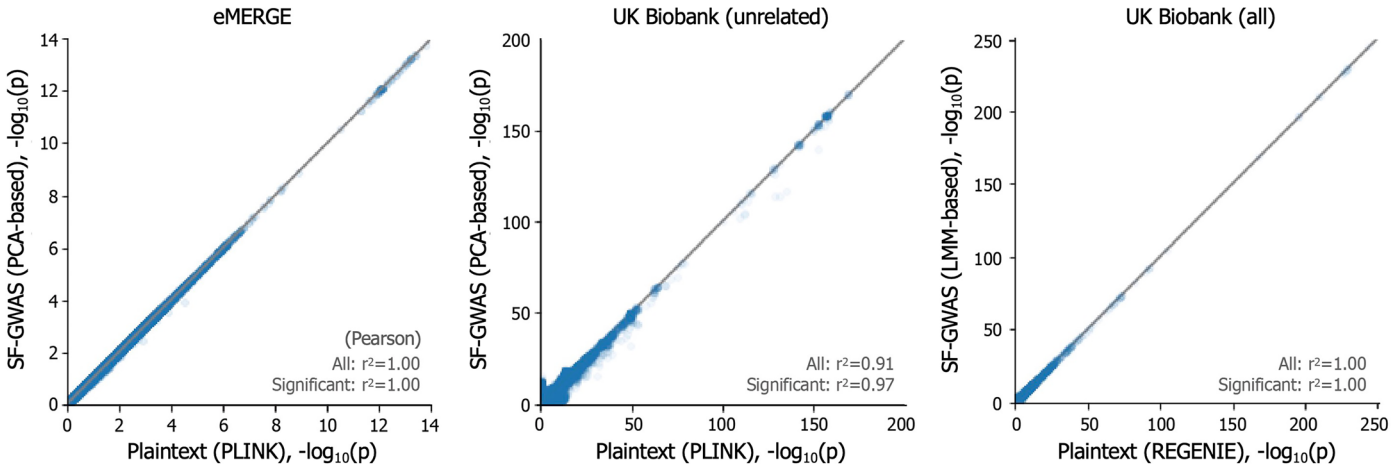
Peer review information *Nature Genetics* thanks Miran Kim and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s41588-025-02109-1



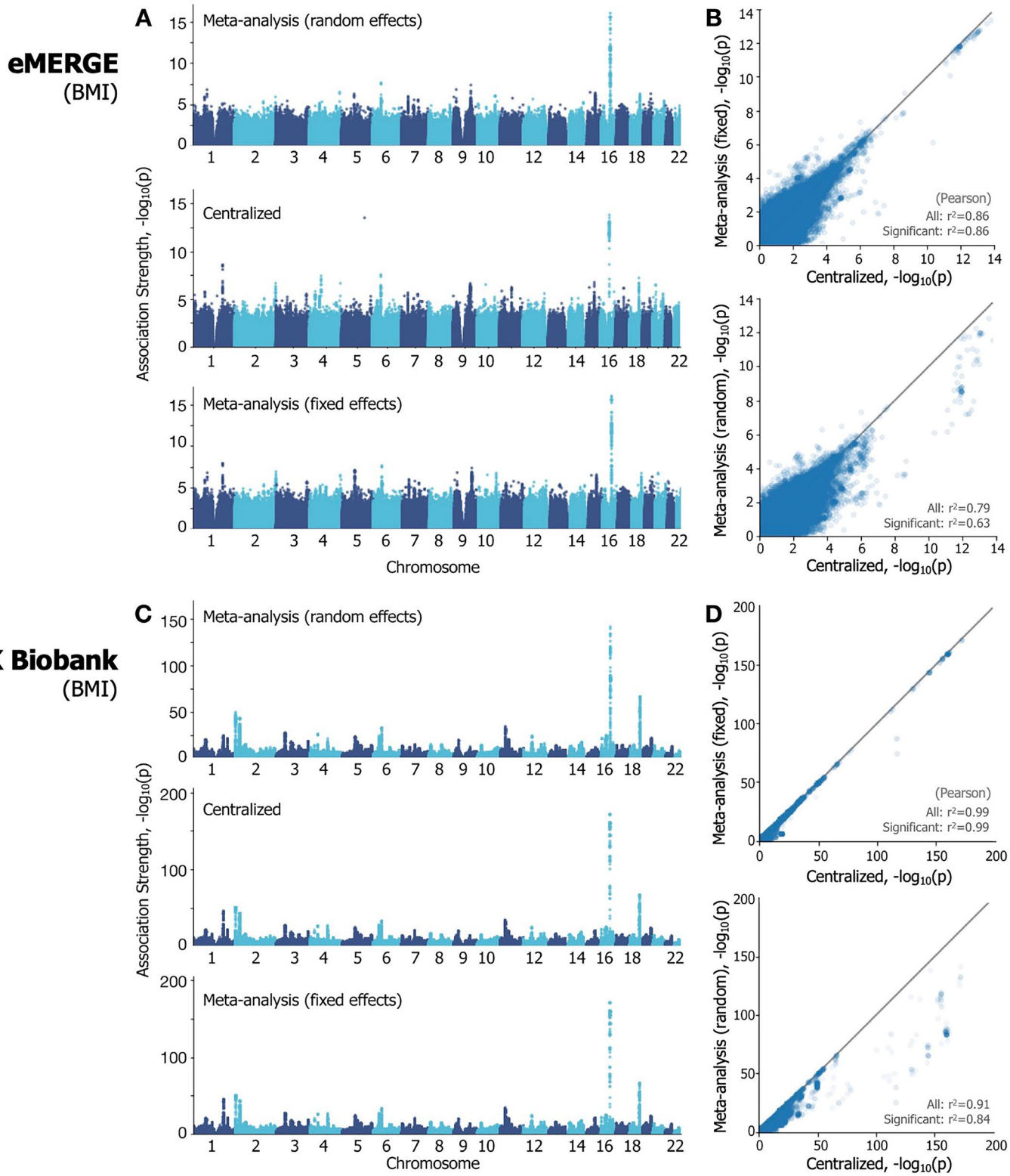
Extended Data Fig. 1 | SF-GWAS accurately reproduces end-to-end PCA-based GWAS (linear) without data centralization. We evaluated PCA-based SF-GWAS (with linear regression) on lung cancer (**left**), bladder cancer (**middle**) and age-related macular degeneration (AMD; **right**) datasets to demonstrate that it obtains similar results as a centralized study in which all plaintext (unencrypted) data are pooled and analyzed together. We evenly split the datasets between two computing parties for joint analysis and compared the association statistics ($-\log_{10}(p)$) of individual genetic variants obtained using (SF-GWAS) with those from a centralized analysis (Plaintext; using the standard scikit-learn Python library). The p-values are based on the two-sided Wald test and are not corrected for multiple testing. The analysis pipeline includes both the quality control and PCA steps. Transparency is added to visualize density. While a small amount of accuracy loss can be seen due to the limited precision of cryptographic schemes, note that the squared Pearson correlation coefficients (r^2) between the two results are always above 0.98. Results for the same datasets using the logistic regression-based tests are provided in Extended Data Fig. 9.

https://doi.org/10.1038/s41588-025-02109-1



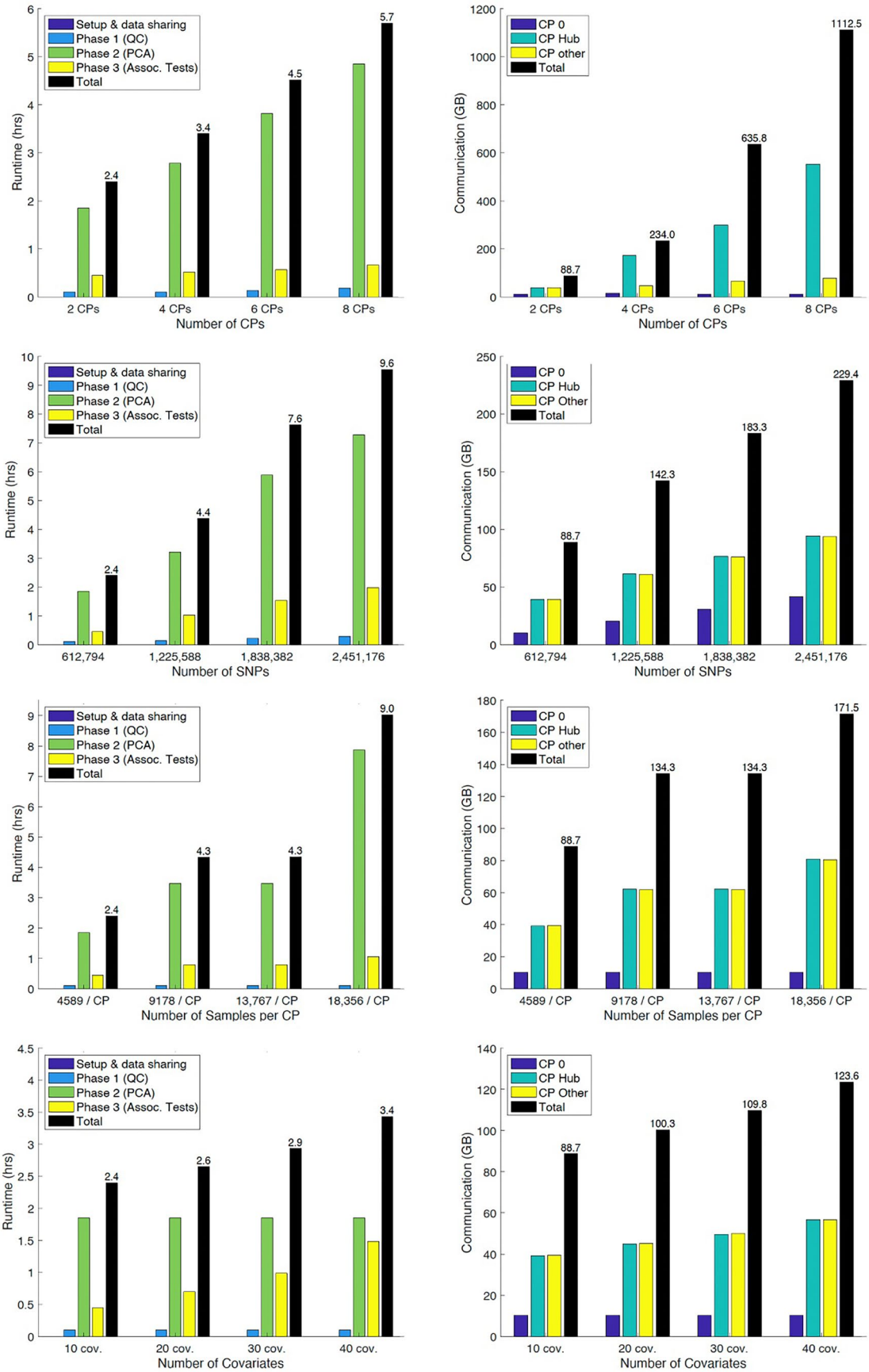
Extended Data Fig. 2 | SF-GWAS accurately reproduces biobank-scale GWAS without data centralization. We evaluated PCA-based SF-GWAS on the eMERGE consortium dataset (**left**), split across 7 data collection centers, and the UK Biobank (UKB) dataset (individuals of European descent, related individuals excluded) (**middle**), split across 6 geographically grouped centers. We additionally evaluated LMM-based SF-GWAS on the UKB dataset including all European individuals, also split across 6 centers (**right**). The plots compare the association statistics ($-\log_{10}(p)$) of individual genetic variants for body mass index (BMI) obtained by SF-GWAS to those from a centralized analysis using the PLINK software for the PCA-based analysis and REGENIE for the LMM-based analysis (Plaintext). The p-values shown are based on the two-sided Wald test for the former and the LMM-based association test for the latter; both are unadjusted for multiple testing. We show the squared Pearson correlation coefficients (r^2) both for all variants and for significant variants identified in the centralized setting (with nominal $p < 5 \times 10^{-8}$). Transparency is added to visualize density. A small amount of numerical noise is attributed to the reduced precision of cryptographic operations. In all datasets, SF-GWAS results closely match the centralized analysis without requiring data centralization.

https://doi.org/10.1038/s41588-025-02109-1



Extended Data Fig. 3 | Meta-analysis approaches for multisite GWAS can deviate from an ideal centralized analysis. Using the same federated study settings from the evaluation of SF-GWAS, we applied standard meta-analysis approaches (that is, random-effect and fixed-effect methods provided by PLINK) to the eMERGE (**top**) and UKB (**bottom**) datasets based on the PCA-based analysis. We show both the Manhattan plots for individual methods (**A, C**) and the scatter plots comparing the association statistics ($-\log_{10}(p)$) between the centralized analysis and each meta-analysis method (**B, D**). The initial p-values in both settings are based on the two-sided Wald test and are not corrected for multiple testing. Squared Pearson correlation coefficients (r^2) both for all variants and for significant variants identified in the centralized setting (with nominal $p < 5 \times 10^{-8}$) are reported. Transparency is added to visualize density in the scatter plots. While meta-analysis results are accurate in UKB based on the fixed-effects model, in all other settings it leads to considerable deviations from the centralized analysis.

https://doi.org/10.1038/s41588-025-02109-1

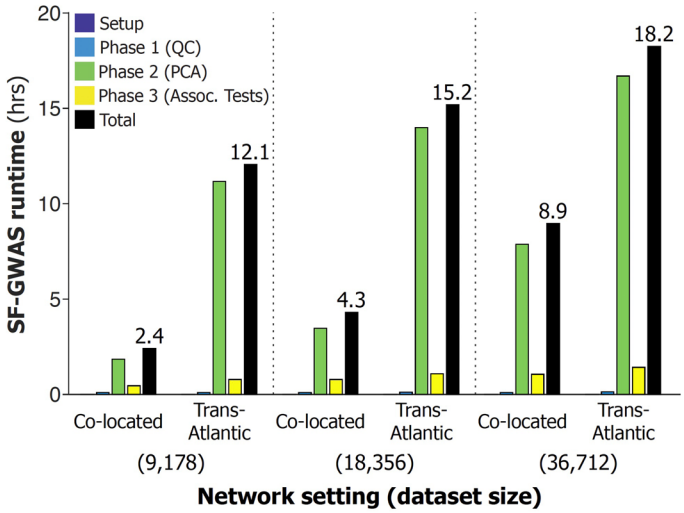


Extended Data Fig. 4 | See next page for caption.

https://doi.org/10.1038/s41588-025-02109-1

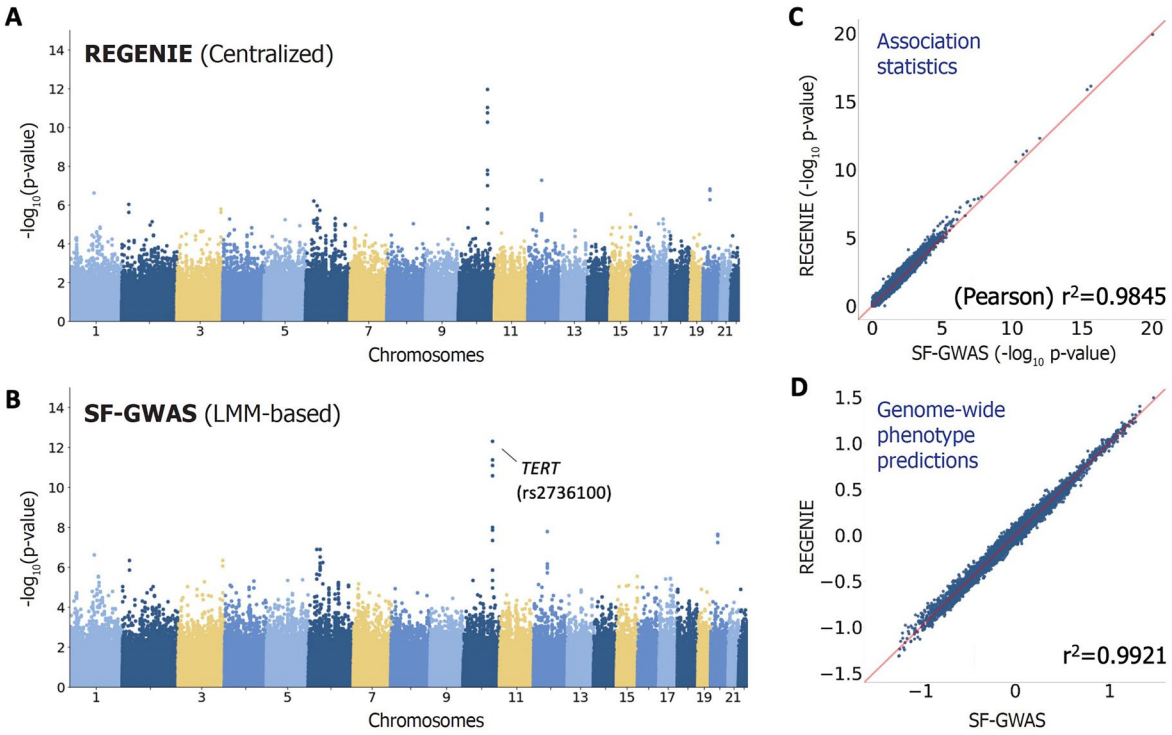
Extended Data Fig. 4 | SF-GWAS efficiently scales with respect to the number of parties, variants, samples, and covariates. We evaluated how the runtime (**left column**) and communication cost (**right column**) of PCA-based SF-GWAS scale with various parameters of the study setting, including: the number of computing parties (CPs) (**first row**), the number of genetic variants (or SNPs; **second row**), the number of covariates (**third row**), and the number of individuals (or samples) per party (**fourth row**). For all experiments, we evenly split the lung cancer dataset between two CPs and replicated as needed to obtain a dataset of desired dimensions. We modified the GWAS parameters to ensure that the amount of data at each step grew proportionally with the input dimensions. For communication, we separately measured the amount of data sent by the auxiliary party in MPC (CP-0); by the "hub" party that aggregates/broadcasts intermediate results (CP-Hub); and by each of the other CPs (CP-Other). The total communication (in black) accounts for data transfer among all parties. Both runtime and communication scale linearly with each parameter. Due to the vector-wise encryption (in groups of 8,192 in our setting), the computational cost increases in steps as the dataset grows (for example, see the identical costs of 9,178 and 13,767 samples per party).

https://doi.org/10.1038/s41588-025-02109-1



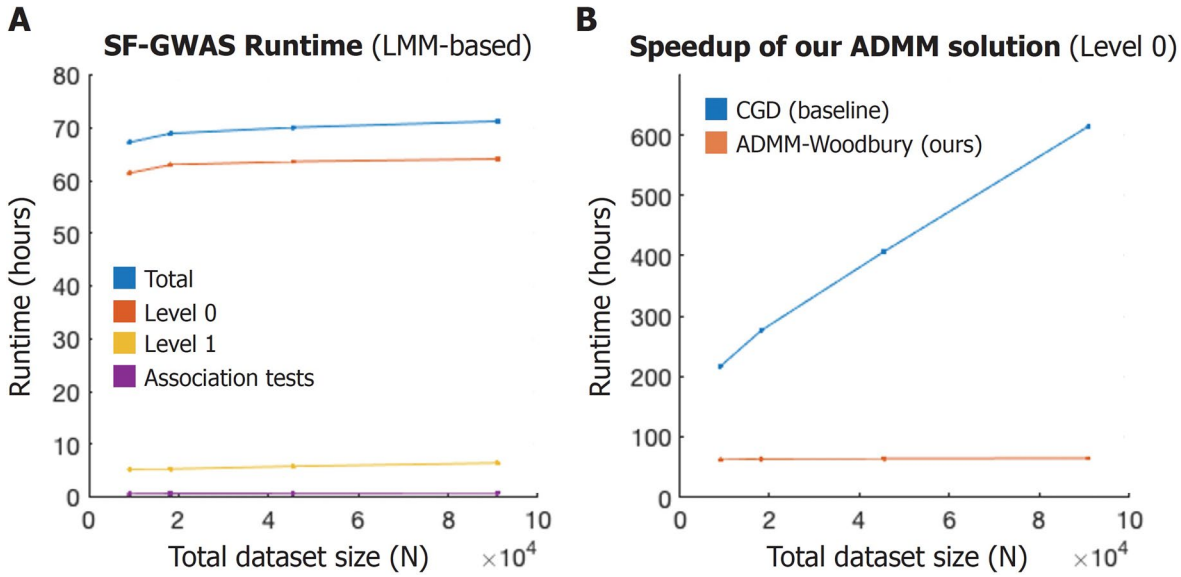
Extended Data Fig. 5 | SF-GWAS remains practical for a trans-Atlantic wide-area network setting between the US and the UK. We evaluated the applicability of SF-GWAS for international collaborations by running the PCA-based workflow on the lung cancer dataset with three computing parties (CPs) located in different geographic regions: two main CPs (each with half of the data) are placed in Oregon (US; "us-central1" in Google Cloud Platform) and London (UK; "europe-west2"), and the auxiliary party (CP-0) is placed in North Virginia (US; "us-east4"). This wide-area network setting (Trans-Atlantic) is compared to the original setting (Co-located) with all CPs in North Virginia (US; "us-east4"). We also replicated the dataset to evaluate the runtime for larger dataset sizes. We observe that SF-GWAS runtime increases by at most 5x when executed among distant machines, and this gap decreases as the dataset gets larger given the increased burden of local computation relative to the communication costs. For a large dataset with 36,712 samples, the trans-Atlantic setting incurs only a 2x slow down. We note that the observed differences are much smaller than the 475-fold increase in the underlying round-trip communication delay in the trans-Atlantic setting; the round-trip latency is 0.2ms between co-located CPs ("us-east4"), 25ms between different regions within the US ("us-central1" and "us-east4"), and 95ms between the US ("us-central1") and the UK ("europe-west2").

https://doi.org/10.1038/s41588-025-02109-1



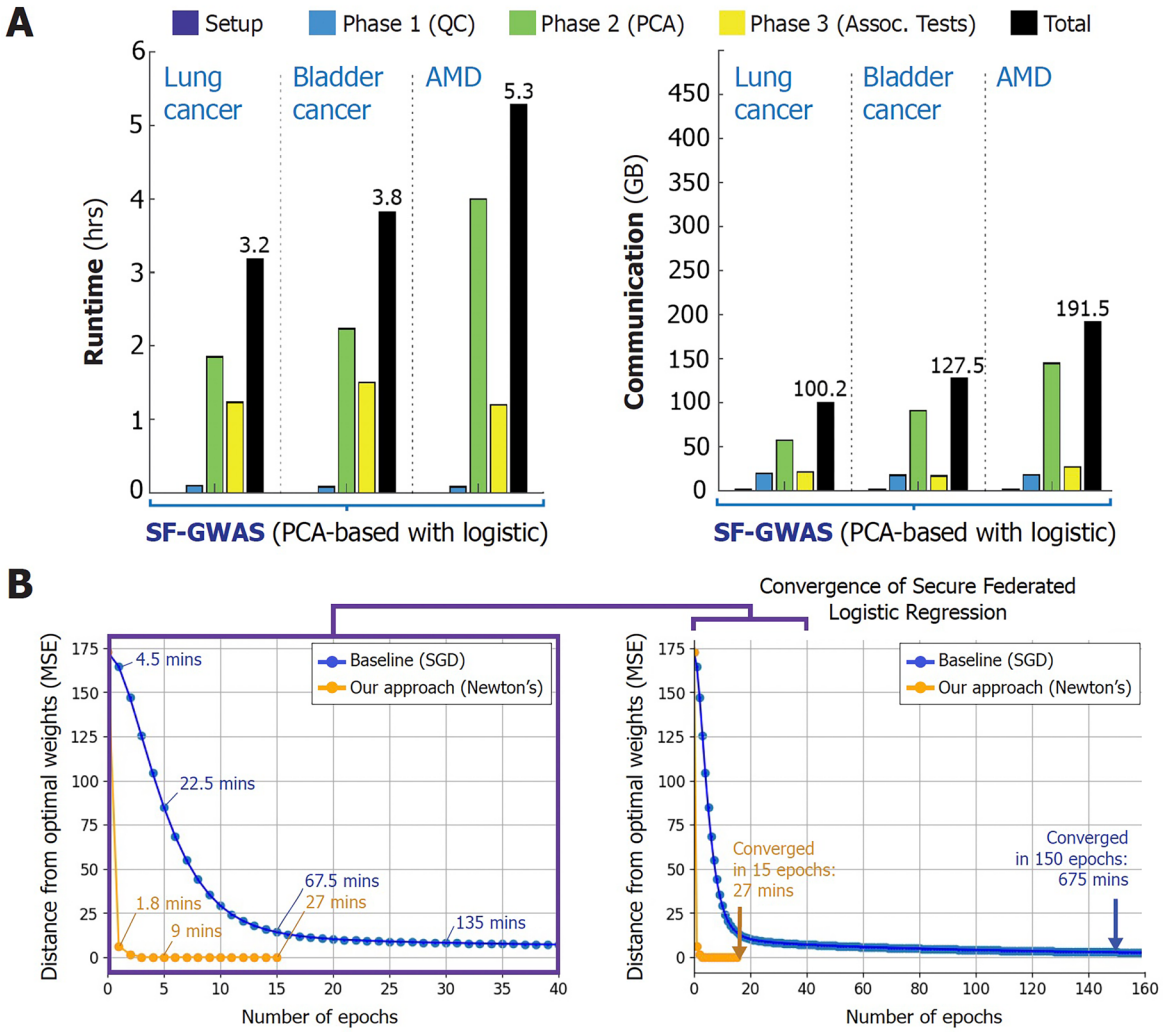
Extended Data Fig. 6 | SF-GWAS closely reproduces REGENIE's LMM-based association tests on the lung cancer dataset. We evaluated LMM-based SF-GWAS on the lung cancer GWAS dataset (including 9,098 individuals and 378,482 SNPs after quality control filters) split between two parties. The Manhattan plot for SF-GWAS (**B**) closely resembles that obtained by running REGENIE on the same centralized dataset (**A**). The p-values shown are based on the LMM-based association test of REGENIE and are reported as two-sided and unadjusted for multiple testing. A particularly strong association is identified for SNP rs2736100, which is associated with the *TERT* gene, a known cancer gene involved in telomere maintenance. (**C**) Comparison of the negative log p-values for all variants in the dataset generated by REGENIE and by SF-GWAS. (**D**) A plot of the genome-wide phenotype prediction vectors obtained by both methods, which capture a portion of phenotypic variation that can be explained by genetic relatedness among individuals. These values are provided as input to the association testing pipeline in the REGENIE algorithm to correct for confounding effects. The results from SF-GWAS closely match those of REGENIE with Pearson correlation coefficients (r^2) greater than 0.98 in both plots.

https://doi.org/10.1038/s41588-025-02109-1



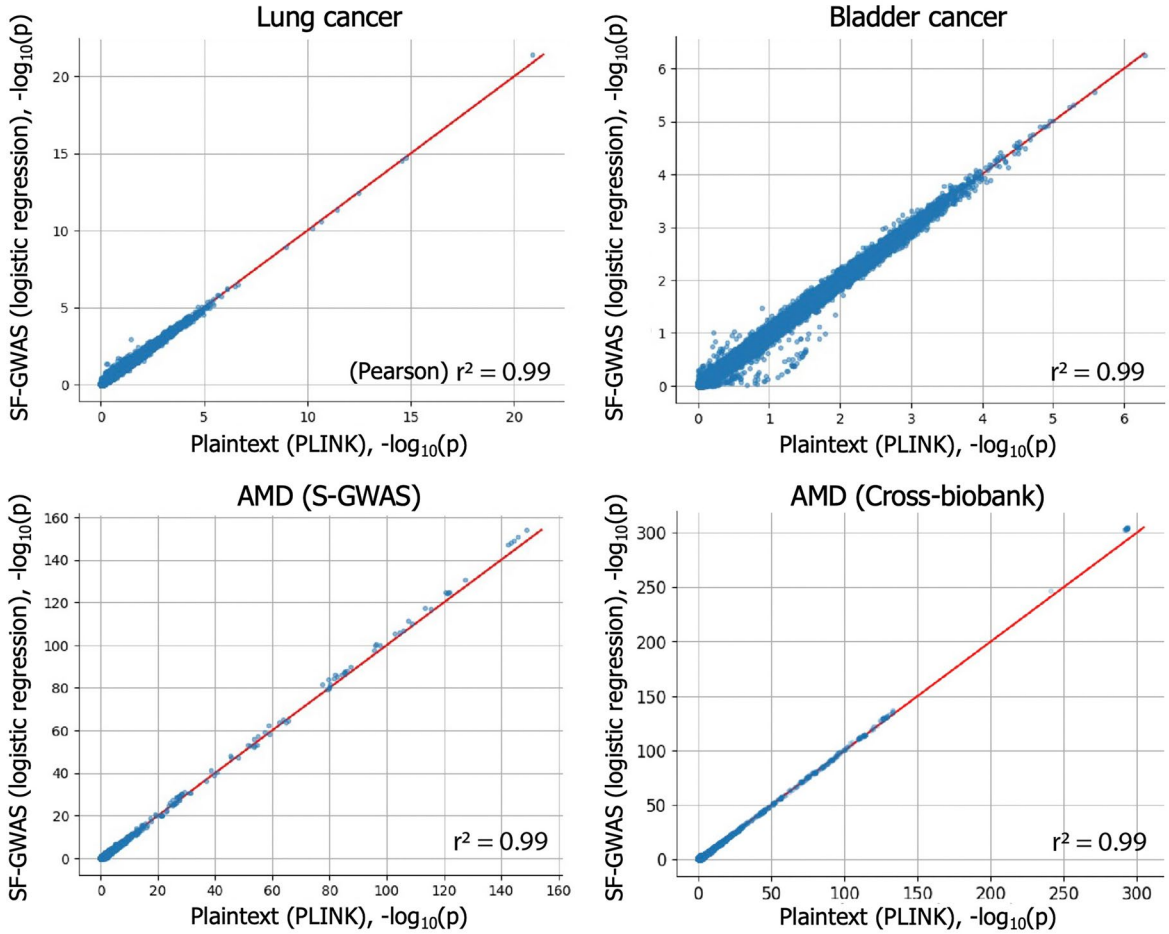
Extended Data Fig. 7 | LMM-based SF-GWAS efficiently scales to large datasets, maintaining a near-constant runtime across varying dataset sizes. (**A**) We show the runtimes of LMM-based SF-GWAS on upsampled lung cancer datasets including up to 91K individuals (ten times the original dataset). Measurements are based on a network of three co-located machines on Google Cloud with 12 cores each for parallelization. Runtimes for iterative components of the algorithm with identical computational load for every iteration are estimated based on a smaller set of iterations. SF-GWAS runtime remains near-constant as the size of the dataset grows due to our scalable design of the federated algorithm. (**B**) We compare our optimized ADMM-Woodbury algorithm for Level 0 of the REGENIE workflow (which is the main computational bottleneck) with the baseline conjugate gradient descent (CGD) algorithm, both implemented in our secure and federated framework. The comparison shows the improved asymptotic complexity of our approach (Supplementary Note). For the largest dataset including 91K individuals, our approach achieves a 9.6-fold speedup over the baseline.

https://doi.org/10.1038/s41588-025-02109-1



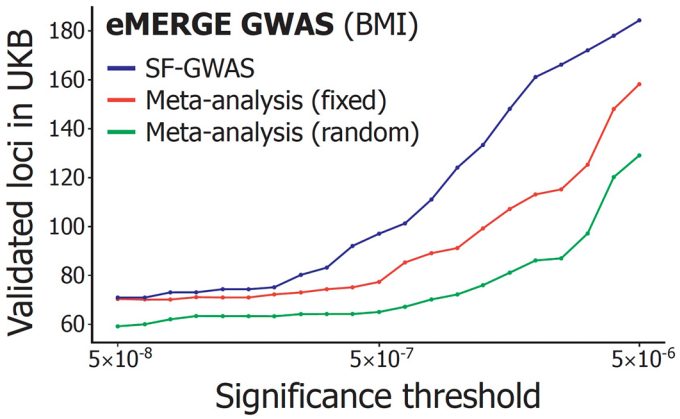
Extended Data Fig. 8 | Logistic regression-based SF-GWAS achieves computational efficiency through fast convergence during model estimation. (**A**) We repeat the performance evaluation presented in Fig. 2 for the logistic regression-based SF-GWAS. Both runtime and communication costs remain comparable to the linear regression-based workflow despite the greater complexity of association testing based on logistic regression models. (**B**) To illustrate the effectiveness of our Newton's method-based solution, we compare the convergence of our secure federated logistic regression pipeline to that of standard mini-batch stochastic gradient descent (SGD)-based optimization adopted by previous work (for example, S-GWAS). We evaluate both implementations on the S-GWAS lung cancer dataset, which includes 9,178 samples and 15 covariates. The dataset is split between 2 parties, with each party using a 16-core machine to perform the computation, analogous to our default experimental setup. The ground truth weights used to evaluate convergence are obtained by fitting the model on a centralized, non-encrypted dataset using standard Python libraries. We note that a range of alternative solvers provided by the "statsmodels" and "scikit-learn" libraries all converged to the nearly identical final weights. We implemented a secure and federated version of SGD with momentum and regularization to compare against our secure and federated approach for Newton's Method. For SGD, we empirically found that a learning rate (alpha) of 0.9, a momentum parameter (mu) of 0.99, a batch size (b) of 200, and a regularization parameter (rho) of 0 (that is, no regularization) were the most effective choices that ensure model convergence in fewer than 200 epochs across all our datasets. In contrast, our Newton's method-based algorithm consistently converged to the optimum within 15 epochs on all datasets. Moreover, our method achieves faster performance in each epoch than SGD (for example, 1.8 vs 4.5 minutes on the lung cancer dataset). This improved efficiency can be attributed to the greater number of iterations required by SGD in each epoch, which are more costly to perform in the secure setting due to the overhead of cryptographic operations. MSE: mean squared error.

https://doi.org/10.1038/s41588-025-02109-1



Extended Data Fig. 9 | SF-GWAS accurately reproduces end-to-end PCA-based GWAS (logistic) without data centralization. We reanalyzed the three case-control GWAS datasets presented in Extended Data Fig. 1 from the S-GWAS publication (lung cancer, bladder cancer, and AMD) using logistic regression-based association tests provided by SF-GWAS. The resulting association statistics ($-\log_{10}(p)$) of individual variants are compared to those of PLINK's logistic regression-based GWAS pipeline. Our p-values are based on the two-sided score-based test for the logistic model and are unadjusted for multiple testing. Squared Pearson correlation coefficients (r^2) between the SF-GWAS and plaintext results are also reported. The same comparison is shown (bottom-right) for our cross-biobank AMD GWAS experiment, where we jointly analyzed three independently collected AMD GWAS datasets from IAMDGC, eMERGE, and UK Biobank (see Methods). SF-GWAS obtains logistic regression-based association statistics that closely match a centralized study in which all plaintext (unencrypted) data are pooled and analyzed together using a standard tool (PLINK).

https://doi.org/10.1038/s41588-025-02109-1



Extended Data Fig. 10 | Collaborative GWAS on eMERGE data using SF-GWAS identifies a greater number of validated associations compared to traditional meta-analysis approaches. We compare SF-GWAS to both fixed-effects and random-effects models of meta-analysis implemented in the PLINK software for a GWAS of body mass index on the eMERGE data (split across 7 data collection centers) with respect to each method's ability to identify validated associations in the larger UK Biobank cohort. We counted an association as validated if the same locus received a significant p-value ($< 5 \times 10^{-8}$) in the summary statistics reported by the Pan-UK Biobank project resource in any ancestry-specific or aggregate analysis of the same phenotype. Plot shows the comparison of the number of validated loci for each method based on varying significance thresholds for the eMERGE analysis from 5×10^{-8} to 5×10^{-6}. Overall, SF-GWAS identified a greater number of validated associations, indicating an increase in statistical power.

nature portfolio

Corresponding author(s): Hyunghoon Cho, Jean-Pierre Hubaux, Bonnie Berger

Last updated by author(s): Nov 20, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	For data processing and baseline association analysis, we used PLINK v2.0 (https://www.cog-genomics.org/plink/2.0/) and the scikit-learn package v1.3.0 (with Python 3.8) for PCA-based pipeline and REGENIE v2.2.4 (https://rgcgithub.github.io/regenie/) for LMM-based pipeline. Michigan Imputation Server v1 used Minimac4 for imputation, which we utilized to obtain imputed genotypes for AMD analysis. GrafPop 1.0 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/Software.cgi) was used for ancestry group assignment. Our code for secure and federated GWAS is available as open source software (v0.1.2) at: https://github.com/hhcho/sfgwas (DOI: 10.5281/zenodo.14726447).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The three datasets used for comparison with the prior work on Secure GWAS (Cho et al., Nature Biotechnology, 2018) are available via NIH dbGaP with accession numbers phs000716.v1.p1 (lung cancer), phs000346.v2.p2 (bladder cancer), and phs001039.v1.p1 (AMD). The eMERGE consortium data is also available via dbGaP (phs000888.v1.p1). Data access applications for the UK Biobank data can be submitted at: https://www.ukbiobank.ac.uk/.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	With the exception of the lung cancer dataset, all datasets in our study includes individuals of both sexes. Where available, we included self-reported sex as a covariate in the association analysis. The lung cancer dataset is obtained from a targeted study of never-smoking East Asian women and includes only females. Sharing of individual-level data for all our datasets is consented by the participants through the informed consent procedure of each respective study.
Reporting on race, ethnicity, or other socially relevant groupings	All our experiments analyze cohorts that were recruited by the original studies. For a subset of experiments considering only individuals of European ancestry, the ancestry label was determined based on self-reported ethnicity for UK Biobank and GrafPop-inferred ancestry for eMERGE and IAMDGC (P_e > 90%). Confounding due to ancestry backgrounds in our analysis is controlled by either including genotype principal components or using linear mixed models.
Population characteristics	Lung cancer dataset: never-smoking East Asian women in all age groups, including 5,088 lung cancer cases and 4,090 controls, recruited by Female Lung Cancer Consortium in Asia (FLCCA). Bladder cancer dataset: individuals of European descent, including both sexes and all age groups, 6,211 cases with histologically confirmed primary carcinoma of the urinary bladder and 6,849 controls, pooled from 16 studies conducted in Spain, Finland, and USA. AMD dataset: 9,648 cases with geographic atrophy (GA), choroidal neovascularization (CNV), or mixed GA/CNV and 13,035 controls, recruited by International Age-Related Macular Degeneration Genomics Consortium, including both sexes and all age groups. eMERGE dataset: a prospective cohort including individuals with electronic medical records (EMRs) and linked genomic data, recruited by individual participating sites in the eMERGE Network (both sexes, aged 3-75). UK Biobank dataset: a prospective cohort of individuals in the UK with EMR and linked genomic data, recruited by the UK Biobank (both sexes, aged 40-69).
Recruitment	See above.
Ethics oversight	Access to all datasets were approved by respective data access committees through NIH dbGaP and UK Biobank Access Management System.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculation was performed to predetermine the sample sizes as we used the largest available public datasets and did not generate new data. We chose a variety of datasets to demonstrate our methods on a range of sample sizes.
Data exclusions	We adopted the following standard quality control filters: genotype missing rate per SNP <0.1, minor allele frequency (MAF) >0.05, and Hardy–Weinberg equilibrium chi-squared test statistic <23.928 (p-value >10^{-6}). We used the same set of filters for the UKB and cross-biobank AMD data, except for MAF >0.001 to account for the larger size of these datasets. These thresholds were established prior to the experiments.
Replication	We confirmed that our experiments can be reproduced based on the code we provide.
Randomization	This is not relevant to our study, as our analyses do not involve distinct experimental groups. Instead, for each dataset, a single analysis is conducted encompassing all samples.

Blinding	This is not relevant to our study, as our analyses do not involve distinct experimental groups. Instead, for each dataset, a single analysis is conducted encompassing all samples.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*