Secure large-scale genome-wide association studies using homomorphic encryption

Marcelo Blatt[a,1], Alexander Gusev[a,b,1], Yuriy Polyakov[a,1,2], and Shafi Goldwasser[a,c,1,2]

[a]Duality Technologies, Inc., Newark, NJ 07103; [b]Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02215; and [c]Simons Institute for the Theory of Computing, University of California, Berkeley, CA 94720

Contributed by Shafi Goldwasser, February 15, 2020 (sent for review October 18, 2019; reviewed by Jung Hee Cheon and David J. Wu)

Genome-wide association studies (GWASs) seek to identify genetic variants associated with a trait, and have been a powerful approach for understanding complex diseases. A critical challenge for GWASs has been the dependence on individual-level data that typically have strict privacy requirements, creating an urgent need for methods that preserve the individual-level privacy of participants. Here, we present a privacy-preserving framework based on several advances in homomorphic encryption and demonstrate that it can perform an accurate GWAS analysis for a real dataset of more than 25,000 individuals, keeping all individual data encrypted and requiring no user interactions. Our extrapolations show that it can evaluate GWASs of 100,000 individuals and 500,000 single-nucleotide polymorphisms (SNPs) in 5.6 h on a single server node (or in 11 min on 31 server nodes running in parallel). Our performance results are more than one order of magnitude faster than prior state-of-the-art results using secure multiparty computation, which requires continuous user interactions, with the accuracy of both solutions being similar. Our homomorphic encryption advances can also be applied to other domains where large-scale statistical analyses over encrypted data are needed.

genome-wide association studies | encrypted computing | homomorphic encryption

A association study (GWAS) evaluates one single-nucleotide polymorphism (SNP) at a time for association to a phenotype or outcome. In the disease case/control setting, this is typically performed through a goodness-of-fit test or logistic regression, which report association odds ratios, standard errors, and *P* values. The results from a GWAS have two broad downstream uses: First, variants that pass a statistical threshold are reported as genome-wide significant and evaluated for functional mechanisms; second, all variants can be integrated into polygenic risk score analyses to predict phenotypes in held-out samples.

A critical challenge for GWASs is the dependence on individual-level data that typically have strict privacy requirements, creating an urgent need for methods that preserve the individual-level privacy of participants (1, 2). There are two main approaches to privacy-preserving GWASs: secure multiparty computation (MPC) and homomorphic encryption (HE). The MPC approach typically uses a protocol invented by Yao in the 1980s called the garbled circuit solution (3, 4). In this protocol, two clouds, each owned by a different hospital or corresponding to two noncollaborating servers within one hospital, hold part of the genomic data to be analyzed. The alternative approach is based on fully HE (FHE), a novel secure encryption method developed in 2008 by Gentry (5), which is much less communication intensive, and is secure even if the servers collaborate. HE allows performing secure computations over encrypted sensitive data without ever decrypting them.

Recent work has focused on secure MPC solutions to facilitate individual-level privacy-preserving GWASs (3, 6). The work of Jagadeesh et al. (3) addressed diagnosis of monogenic diseases while preserving participant privacy using MPC. Due to the communication and computationally intensive nature of the garbled circuit solution, GWASs beyond monogenic diseases were not addressed, and the patient cohort was small. Jagadeesh et al. estimated that, even for the monogenic example, garbled circuits would be at least 5,000 times faster than FHE. Cho et al. (6) followed by successfully computing a GWAS by dividing data among multiple servers and computing the GWAS via multiparty secure protocol among the servers, subsets of which are trusted not to collaborate against other servers, else privacy is lost. Here, we no longer need to resort to this trust assumption. We are successfully using HE to encrypt the genomic sequences of study participants while enabling GWAS computations without the ability to decrypt, and scaling to hundreds of thousands of samples (Fig. 1).

We implement two common GWAS techniques—the allelic chi-square test for case control differences and a logistic regression approximation (LRA) with covariates—within our HE framework. The LRA algorithm utilizes a previously proposed semiparallel approach to efficiently iterate over each genetic variant without requiring repeated likelihood maximizations (7). Our HE LRA implementation of this approach was

Significance

We propose a toolbox of statistical techniques that leverage homomorphic encryption (HE) to perform large-scale GWASs on encrypted genetic/phenotype data noninteractively and without requiring decryption. We reformulated the GWAS tests to fully benefit from encrypted data packing and parallel computation, integrated highly efficient statistical computations, and developed over a dozen cryptoengineering optimizations. Our HE solution is >30× faster than the cutting-edge multiparty computation method, in contrast to the claim that HE is not viable for large-scale GWASs. Our approach is thus a significant advance both in methodology and application, empowering large-scale cross-institution collaboration, patient-driven research, and crowdsourced genomics.

Author contributions: M.B., A.G., Y.P., and S.G. performed research and wrote the paper.

Reviewers: J.H.C., National Seoul University; and D.J.W., University of Virginia.

Competing interest statement: S.G. is the director of the Simons Institute at University of California, Berkeley. She is also a cofounder of Duality Technologies to which she consults 1 day a week. M.B. is the Head of Data Science of Duality Technologies. A.G. performed the initial GWAS work for the LRA and chi-square test solutions (developed as part of the iDASH'18 competition) as a consultant of Duality Technologies, Inc. and did the work specific to the AMD dataset analysis as an Assistant Professor at Dana-Farber Cancer Institute. Y.P. is a Research Scientist at the New Jersey Institute of Technology and is also a Principal Scientist at Duality Technologies.

Data deposition: The implementation of the GWAS protocols developed in this work is publicly available in GitLab at https://gitlab.com/duality-technologies-public/palisade-gwas-demos/.

[1]M.B., A.G., Y.P., and S.G. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: ypolyakov@dualitytech.com or shafi@csail.mit.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1918257117/-/DCSupplemental.

First published May 12, 2020.

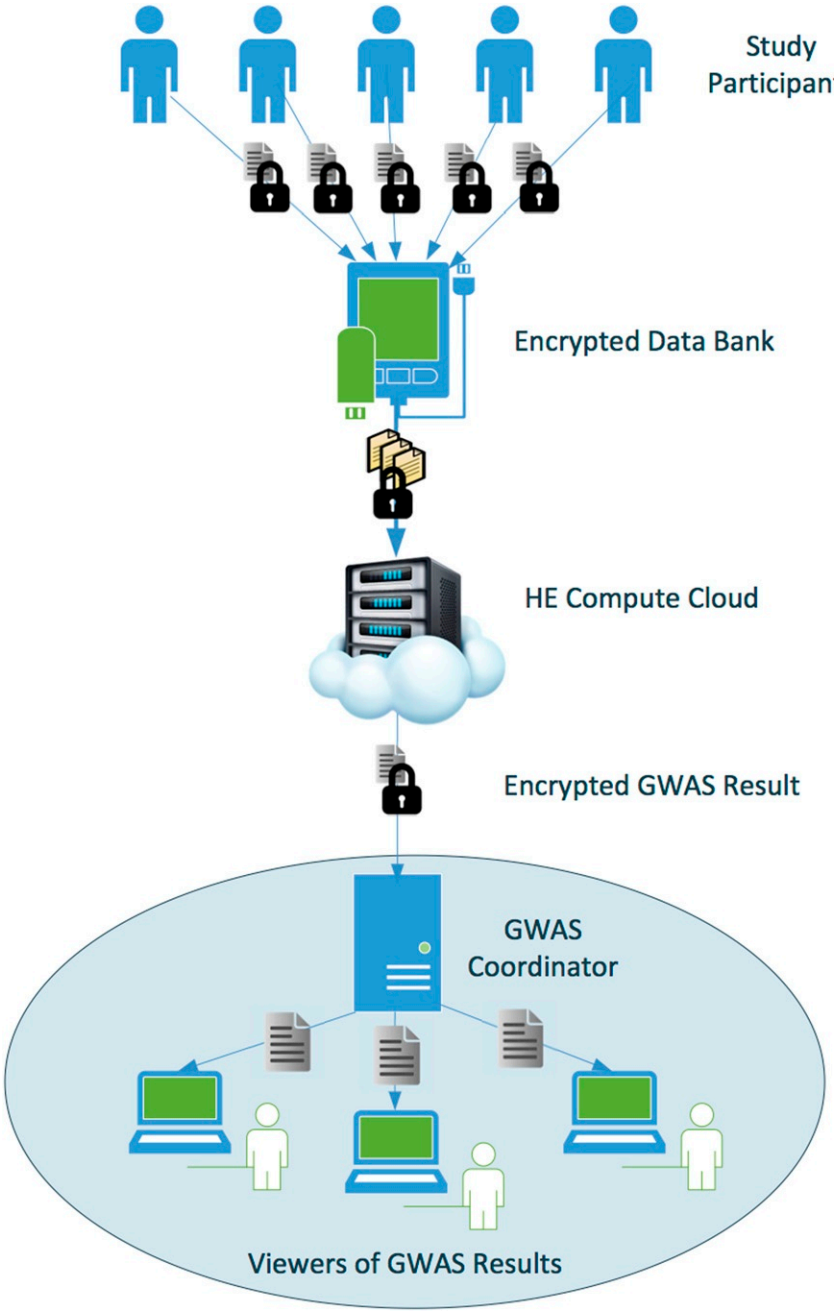

Fig. 1. Schematic of the HE GWAS. First, study participants obtain a public key from the GWAS coordinator (this step is not shown in the figure, for simplicity). Then, each of them encrypts their data using the public key, and sends the encrypted data to the Encrypted Data Bank, storing all encrypted individual-level data from many study participants. When a specific study is initiated by the GWAS coordinator, the encrypted data for the individuals in the study get transmitted to the HE Compute Cloud for a noninteractive secure computation. Next, the HE Compute Cloud computes the results and sends them in encrypted form to the GWAS coordinator. Finally, the GWAS coordinator decrypts the results and routes them to one of the viewers.

independently tested in the iDASH 2018 secure genome analysis competition (http://www.humangenomeprivacy.org/2018/) and received the first place.* We additionally present a highly efficient chi-square test that is faster than the LRA implementation by a factor of 40× and consumes 6× less memory at the cost of excluding covariates from the model. Our HE framework provides postquantum security and is based on several advances. First, we reformulated the compute models for both the chi-square and LRA algorithms to fully benefit from ciphertext packing, enabling the parallel execution of thousands of multiplications/additions using a single homomorphic multiplication/addition. Second, we introduced two types of data encoding to minimize the number of computationally expensive key switching operations, and developed several methods for converting between the encodings homomorphically (used in the LRA solution). Third, we applied multiple plaintext approximations for the LRA model. Fourth, we developed an efficient residue number system (RNS) variant of the

*Our HE LRA solution shared first place with the solution by the team from the University of California San Diego.

Cheon–Kim–Kim–Song (CKKS) HE scheme (8), which naturally supports approximate number arithmetic. Finally, we applied more than a dozen cryptoengineering optimizations.

We apply our HE framework to a real GWAS of age-related macular degeneration (AMD) (9) of 12,461 cases and 14,276 controls (restricted to self-reported Europeans) genotyped on 263,941 total markers with minor allele frequency of >1%. For our gold standard, we computed association statistics in the clear using full logistic regression on each variant with sex, age, and age squared as covariates. We first compared the distributions of GWAS statistics on a subset of 16,384 SNPs and 5,000 individuals evaluated by the same statistical test with/without HE, where we expect essentially perfect concordance (Fig. 2 *A* and *B*). Both the chi-square and LRA tests produced HE statistics with an R^2 of 1.00 to the statistics in the clear, and a replication slope of 1.00 and 0.98, respectively, indicating negligible bias (see *Materials and Methods*). We next compared the HE GWAS statistics to the gold standard logistic regression statistics, with any differences now arising from both the statistical assumptions and the HE (Fig. 2 *C* and *D*). The LRA again produced highly accurate HE statistics, with an R^2 of 1.00 and a replication slope of 0.98. The HE chi-square statistics exhibited some loss of signal relative to the gold standard but remained highly robust, with an R^2 of 0.96 (replication slope 0.99) primarily due to noise at nonsignificant variants. Importantly, the chi-square test odds ratios remained highly accurate and nearly unbiased (R^2 = 0.95; *SI Appendix*, Fig. S1). We confirmed that accuracy was high across all variants by computing polygenic risk scores, wherein genetic risk values are predicted for each individual as the sum of risk alleles they carry weighted by the allelic effect size (see *Materials and Methods*, Fig. 2 *E* and *F*, and *SI Appendix*, Fig. S2). Both risk scores were highly correlated with scores from the gold standard logistic test ($R^2 > 0.99$) with a slight downward bias in the absolute score for the LRA (replication slope 0.95) (Fig. 2 *E* and *F*). Finally, we examined individual genome-wide significant associations reported by the original GWAS, where we again observed highly concordant results with the gold standard using both statistical tests (Table 1).

We next downsampled the data to investigate the run time and scalability characteristics as a function of SNPs and sample size (Fig. 3 and *SI Appendix*, Tables S1–S6). We found the LRA computation to scale linearly in the number of markers and the number of individuals. At the largest evaluated sample size of n = 15,000 and M = 16,384, computation took 1.1 h, extrapolated to 7.7 h for a GWAS of n = 100,000 (M = 16,384), and extrapolated to 234 h for a GWAS of n = 100,000 and M = 500,000 run as a single job. By comparison, the chi-square test (which requires only simple mathematical operations) was more than an order of magnitude faster than the LRA, completing the same analysis of n = 15,000 and M = 16,384 in 98 s (41× faster than the LRA). This was extrapolated to 11 min for n = 100,000 (M = 16,384) or 5.6 h for a GWAS of n = 100,000 and M = 500,000. The chi-square solution also required approximately 6× less peak RAM and thus could be run on a full-scale cohort of n = 25,000 (M = 49,152) in 8 min within the memory constraints available to us (*SI Appendix*, Table S4). Detailed run time characteristics including encryption/decryption are available in *SI Appendix*, Tables S1, S2, S4, and S5. As both the LRA algorithm and chi-square test are natively parallel over the number of SNPs, the computations can be trivially distributed to multiple nodes, with each node working with 16,384 SNPs at a time (see *Materials and Methods*). This implies that a GWAS of n = 100,000 and M = 500,000 could be run in 11 min on 31 nodes running in parallel.

Our HE solution for the chi-square test is faster than the state-of-the-art MPC approach of Cho et al. (6) extrapolated to 100,000 individuals and 500,000 SNPs: 5.6 h (for HE) vs. 37 h (for MPC association tests only, without quality control or population stratification analysis; Phase 3 in figure 2a of ref. 6) or 193 h (for full MPC). The accuracy of both solutions is similar. Our LRA solution has a run time of 234 h for this scenario, while the previously published MPC approach "did not yield a practical runtime for a genome-wide application of logistic regression." Both of our solutions are fully noninteractive, produce valid odds ratios in the analyses of real data, and natively parallelize over the number of SNPs, enabling their execution in distributed computing cloud environments. Although we did not implement it here, a hybrid approach where all SNPs are evaluated with the chi-square test and then the 5% most significant SNPs are retested by the LRA could also be used to achieve the same accuracy as LRA for significant associations, requiring only 17 h (excluding ciphertext repacking overhead, which would be relatively small).

Our approach has several limitations and areas of future work. First, unlike previous work (6), our model assumes that encrypted data have been fully processed and does not perform additional quality control or genetic ancestry inference, although such methods can be easily applied preencryption. In particular, Chen et al. (10) showed that fine-scale genetic ancestry is much more accurately inferred by projection from external population reference data than by principal component analysis directly on the target samples and leads to more effective correction for population stratification. High-quality population reference data are available for all major populations, and the preencryption data can be easily projected using these references to compute ancestry covariates [requiring a simple matrix–vector product, as, for example, implemented in the PLINK score function (11)]. Second, while the chi-square test requires no parameter tuning, the LRA relies on a learning rate parameter (see *Materials and Methods*) that may differ by study depending on size and relationship of covariates. This can be circumvented by tuning the parameter on subsets of the data in the clear, or by comparing to parameter-free solutions such as the chi-square or linear regression results, at the cost of some additional computation. Third, our approach does not prevent the HE Compute Cloud from colluding with the GWAS coordinator to decrypt the original data, which is also true for existing MPC solutions. This problem can be addressed by adding a secret sharing protocol or using a variant of threshold HE (12) described in the next paragraph.

Extensions to a multiparty scenario are possible using threshold HE (12), a protocol where many parties cooperatively generate a common public key using their individual secret keys ("secret shares"). In this setting, the joint secret key corresponding to the common public key is never seen by any party. In GWASs, the same genotypes and phenotypes can be transmitted from multiple participants and then combined together, or genotypes and phenotypes can be separately transmitted for the same individuals from different participants and then joined together. This extension does not add substantial computation overhead to our single-party HE solution (the computation itself is performed the same way). Our work here is thus a step toward enabling analyses of sensitive phenotypes that cannot be shared between groups/institutions and individual patient participation in research studies without risk to genomic privacy.

Many of our HE improvements are general-purpose and can be applied to other application domains where similar large-scale association and regression tools are used, including phenome-wide association studies from electronic medical record data (13), discovery of predictors of treatment response in clinical trials (14), and correlative studies of multimodal data such as expression/microbiome activity (15). The tests developed here can also be extended to richer machine learning models, including decision and gradient boosted trees.

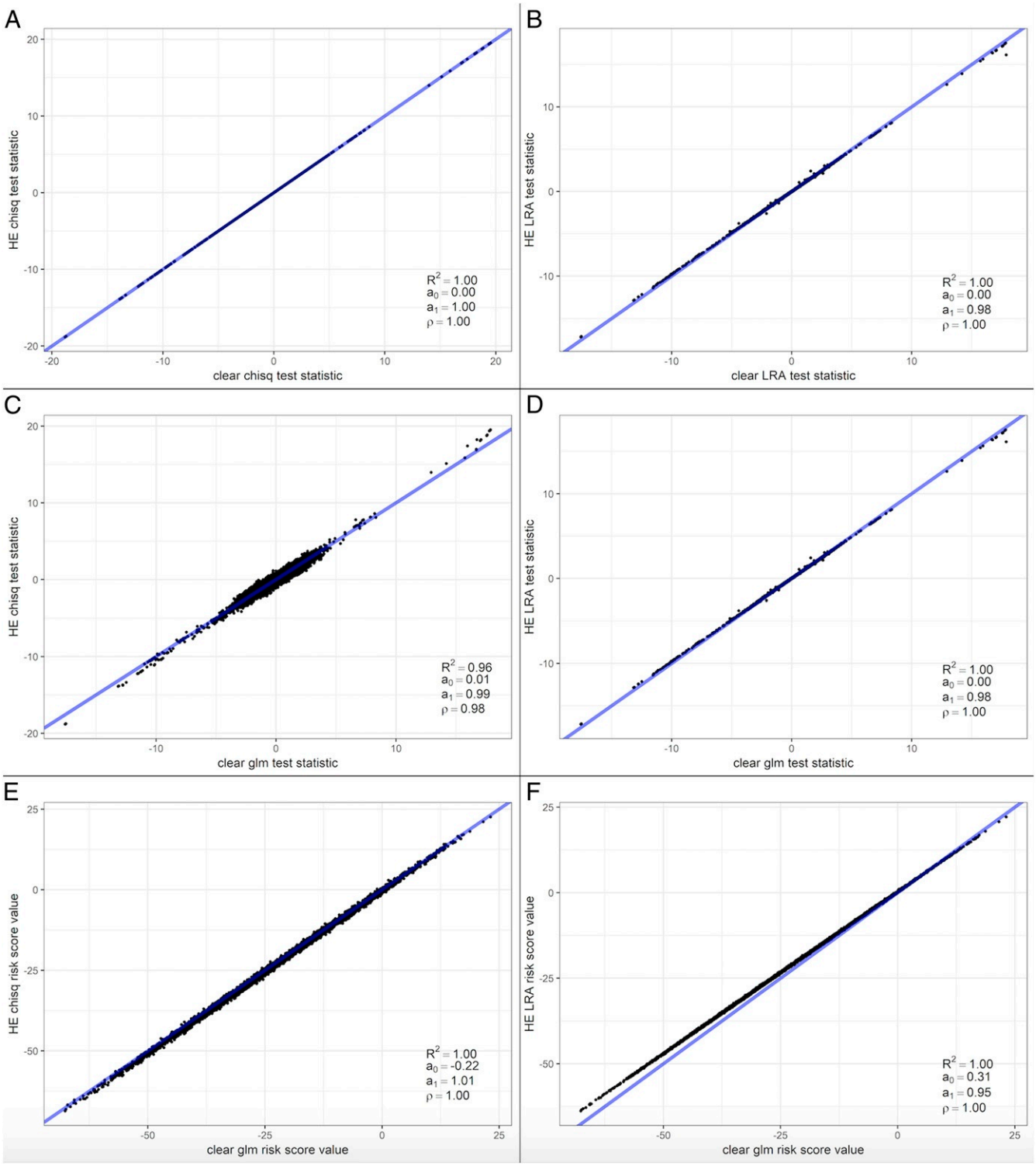

Fig. 2. Highly accurate HE GWAS test statistics and polygenic scores. Each plot shows a signed test statistic computed in the clear (x axis) and the corresponding statistic computed using an HE test (y axis). The "chisq" and "LRA" refer to the chi-square and logistic regression approximation techniques, respectively. *A* and *B* report the same test performed in the clear versus through HE. *C* and *D* report logistic regression (glm test statistic) performed in the clear versus the HE tests. *E* and *F* report polygenic risk scores computed from logistic regression odds ratios (glm risk score value) in the clear versus from the HE tests (restricted to SNPs with association $P < 0.01$). R^2, coefficient of determination of the regression; a^0, intercept of the regression; a^1, slope of the regression; ρ, correlation of statistics.

Materials and Methods

HE. Our solution is based on an optimized variant of the CKKS scheme (8), which is designed for performing approximate number arithmetic homomorphically. We have developed a Double-Chinese Remainder Theorem (CRT), aka RNS, variant of the original scheme. Our variant is based on the same security assumptions as the original scheme, namely, the Ring Learning With Errors (RLWE) problem, but relies on native 64-bit integer arithmetic instead of multiprecision integer arithmetic for better

Table 1. Association statistics from clear and HE tests at known AMD SNPs

	GLM		HE LRA		HE Chisq	
SNP	OR	stat	OR	stat	OR	stat
rs10033900_T	1.09	1.97	1.08	1.91	1.06	1.44
rs943080_C	0.88	−2.94	0.89	−2.88	0.91	−2.26
rs79037040_G	0.88	−2.98	0.88	−2.91	0.89	−2.82
rs2043085_T	0.91	−2.01	0.92	−1.95	0.92	−2.13
rs2230199_C	1.41	6.83	1.38	6.67	1.40	7.10
rs8135665_T	1.12	2.04	1.12	2.03	1.12	2.29
rs114203272_T	0.62	−3.55	0.63	−3.50	0.67	−3.08
rs114212178_T	0.87	−0.70	0.87	−0.69	0.86	−0.77

Reported AMD SNPs were tested for association in a subset of n = 5,000 samples from the AMD study using gold standard logistic regression (GLM), the HE LRA, and the HE chi-square (Chisq) test. "OR" reports the odds ratio; for GLM and LRA, "stat" reports the test statistic; for comparison, the chi-square test "stat" reports the square root of the statistic polarized on the direction of the OR.

performance and parallelization. The RLWE problem is immune to all known classic/quantum computer attacks, and is used as the basis for the HE security standard (16).

The main differences of our Double-CRT variant compared to the original scheme are 1) an efficient rescaling algorithm that works with residues directly, and does not require switching to a slower positional (multiprecision) number system, and 2) an efficient key switching operation previously used for the Brakerski/Fan-Vercauteren scheme (17, 18). This key switching algorithm was originally proposed by Bajard et al. (19) and improved by Halevi et al. (20).

Our variant and parameter selection for the LRA implementation are described in detail by Blatt et al. (21) and also included in *SI Appendix* for completeness.

The CKKS HE scheme has also been extended to an FHE setting (22–24), which supports ciphertext refreshing via bootstrapping when further computations (e.g., after GWAS analysis) need to be performed. Although we did not use bootstrapping in our HE solutions, as the computation circuits for both the LRA and chi-square algorithms are known in advance, our HE framework can be extended to this more general scenario.

Our work differs from our previous work in ref. 21 and the corresponding iDASH analysis in multiple key ways. First, we introduce the highly efficient chi-square test, which is 40× faster and consumes 6× less memory than the LRA proposed in ref. 21. Second, we evaluate the performance and accuracy of both tests using a published GWAS of 26,000 case/controls samples across 260,000 SNPs, whereas the implementation in ref. 21 was evaluated over a toy dataset of 245 case/control samples with 10,643 SNPs (the majority of which were rare variants) and was thus not investigated in a production-level GWAS setting. The accuracy analysis in this manuscript additionally includes accuracy of polygenic risk scores, which were not considered in ref. 21. Third, we evaluate both methods across many data settings and extrapolate performance to 100,000 individuals and 500,000 SNPs, reflecting the scale of emerging GWASs. Fourth, we consider applications, distributed computation, parallelization, and extensions to multiparty scenarios that were not discussed in ref. 21.

Software Implementation. We implemented our solution in PALISADE v1.4.0 (25), an open-source lattice cryptography library. We added our own implementation for the RNS variant of the CKKS scheme to PALISADE (made publicly available in PALISADE starting with v1.7). For loop parallelization, we used OpenMP.

Experimental Test Bed. Experiments were performed using a server computing node with two sockets of Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40 GHz, each with 14 cores; 500 GB of RAM was accessible for the experiments. The node had Fedora 26 OS and g++ (GCC) 7.1.1 installed.

Note that we kept all keys and ciphertexts loaded in the memory to show the total storage requirement for both solutions. In a practical setting, ciphertexts could be serialized to and deserialized from persistent storage, such as solid-state drives, as needed, for example, working with 16,384 SNPs at a time. In this case, the memory requirements would be significantly smaller than in our experiments, and would remain essentially constant when the number of SNPs is increased.

Logistic Regression Approximation. Our LRA solution is based on the semi-parallel method of Sikorska et al. (7). We applied a number of approximations to optimize the HE solution. Our approximations are described in *SI Appendix*. We focused on the case/control setting and thus did not evaluate a standard linear regression, but we note that it is a subproblem of the LRA test that requires less computation and has been previously demonstrated in the HE setting in the iDASH'18 competition by the University of California San Diego team (8).

HE LRA Solution. Our HE LRA solution is described in detail in *SI Appendix*. In summary, we introduced two plaintext encodings and developed several methods for switching between the encodings. We also applied more than a dozen cryptooptimization techniques. The only differences in the HE implementation for the AMD dataset compared to previous work (21) are in the values of the learning rate and auxiliary scaling factors in the HE solution.

The current LRA implementation is limited to three regression covariates, although we believe the method can model up to five covariates relatively efficiently using the same approach (Cramer's rule for matrix inversion) and a greater number of covariates using an approximate technique for matrix inversion discussed by Cheon et al. (26).

Allelic Chi-Square Test. We implemented a standard one-degree-of-freedom allelic chi-square test for difference in major/minor allele counts between cases and controls. Under Hardy–Weinberg equilibrium (enforced here through genotype QC), this test is equivalent to the genotypic (2 × 3) chi-square test (27) or the Cochran Armitage trend test used previously (6). The chi-square HE solution is described in detail in *SI Appendix*.

GWAS Dataset Processing. The GWAS data were downloaded from dbGAP (phs001039.v1.p1) and restricted to all self-identified European samples and QC passing SNPs with minor allele frequency of >1%. The gold standard logistic regression was run using sex, age, and age squared as covariates using the standard glm function in R. The LRA analyses were carried out with the same set of covariates, and the chi-square test analyses were carried out with no covariates.

Accuracy Metrics. We evaluated test accuracy using two metrics: R^2, computed as the coefficient of determination from a regression of the estimated

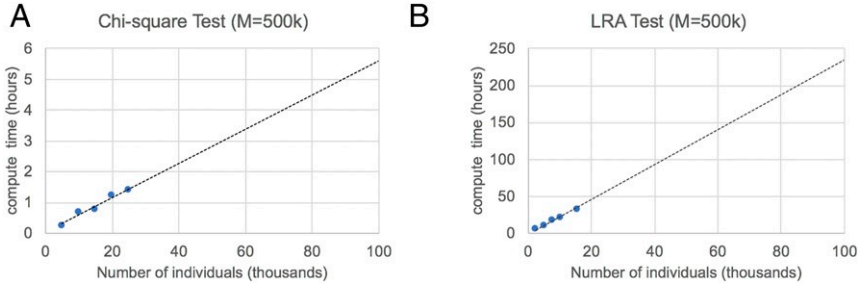

Fig. 3. Linear run time scaling and extrapolation to 100,000 individuals. Run time measured from down-sampling individuals in an analysis of 16,384 SNPs, extrapolated to M = 500,000 SNPs and the given sample size (x axis) using a linear fit. Measured results are shown with points, extrapolated fit with dashed line. (*A*) HE chi-square test; (*B*) HE LRA test.

test statistic on the ground truth, and replication slope, computed as the slope of the regression. The R^2 reflects how much variance in the ground truth statistic is explained by the estimate. The replication slope reflects the scaling factor on the effect sizes imposed by the estimation. When the estimated effect size distribution is linear, the squared replication slope of the test statistics can be thought of as the effective decrease in sample size due to estimation noise (28).

Polygenic Risk Score. We implemented a simple threshold-based polygenic risk score to avoid parameter tuning. After computing the GWAS statistics, variants passing a given P value threshold were retained ($P < 0.01$ or $P <$ 5e-8) and used to predict the genetic value of each individual in the study. The prediction for each sample was the sum across all SNPs of the number of major alleles the individual carries times the major allelic odds ratio of that SNP. We did not account for linkage disequilibrium across markers (i.e., through pruning), because we were only interested in the relative accuracy of the genetic value computed from different tests.

Run Time Extrapolation. For the chi-square test, run time was computed for all n = 25,000 samples in increasing blocks of M = 16,384 SNPs until maximum RAM capacity was reached at M = 49,152, as well as for a single block of M = 16,384 SNPs from n = 5,000 to n = 25,000 in steps of 5,000 (*SI Appendix*, Table S4). For the LRA, which required substantially more RAM, run time was computed for n = 5,000 samples in increasing blocks of M = 16,384 SNPs until maximum RAM capacity was reached at M = 65,536, as well as a single block of M = 16,384 from n = 2,500 to n = 15,000 in steps of 2,500. A linear trend line was then fit to the subsampled data to extrapolate to larger SNP/sample sizes; the linear fit was highly accurate, producing an $R^2 > 0.98$ for both tests (*SI Appendix*, Tables S3 and S6). Linear extrapolation was similarly used in previous published work (6).

Memory Extrapolation. We measured peak RAM usage (i.e., the total storage requirement) after downsampling SNPs at a fixed sample size (*SI Appendix*, Table S7), or individuals fixed at 16,384 SNPs (*SI Appendix*, Table S8). Extrapolation was then calculated from downsampled individuals using a linear fit, which was highly accurate ($R^2 > 0.99$) (*SI Appendix*, Table S8). We note that computations involving individuals that cannot be fully stored in memory (e.g., millions) can be computed in large individual subsets and merged by metaanalysis with negligible loss of accuracy, as is typically done for large-scale GWASs involving multiple consortia.

Distributed Computation and Parallelization. Both LRA and chi-square test algorithms perform computations for each SNP independently. Our implementations use ciphertext packing and hence perform GWAS computations for batches of 16,384 and 4,096 SNPs at a time for LRA and chi-square test, respectively. This implies that the GWAS computation for a large number of SNPs can be trivially distributed to multiple nodes by sending different batches to different nodes in parallel. For instance, we can securely evaluate a GWAS of n = 100,000 and M = 500,000 using the chi-square test in 11 min on 31 nodes if batches of 16,384 SNPs are sent to different nodes in parallel, vs. 5.6 h when a single node is used for the whole computation.

Data Availability. Our analysis is based on the phs001039.v1.p1 dataset available for download in dbGAP. The pseudocode for chi-square test and LRA HE protocols is listed in *SI Appendix*, algorithms 2 and 5, respectively. The implementation of all cryptographic capabilities used in our work, including our optimized CKKS variant, is publicly available for download in PALISADE v1.7.4 and later (25). The implementation of the GWAS protocols developed in this work is publicly available (29).

ACKNOWLEDGMENTS. We gratefully acknowledge the input and feedback from Kurt Rohloff and Vinod Vaikuntanathan. We also acknowledge Kurt Rohloff for the funding of the initial iDASH'18 work. Research reported in this publication was supported, in part, by National Human Genome Research Institute of NIH under Award 1R43HG010123. A.G. was supported by NIH Grant R01CA227237 and the Claudia Adams Barr Award.

1. M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, Y. Erlich, Identifying personal genomes by surname inference. *Science* **339**, 321–324 (2013).
2. S. E. Brenner, Be prepared for the big genome leak. *Nature* **498**, 139–139 (2013).
3. K. A. Jagadeesh, D. J. Wu, J. A. Birgmeier, D. Boneh, G. Bejerano, Deriving genomic diagnoses without revealing patient genomes. *Science* **357**, 692–695 (2017).
4. A. C.-C. Ya, "How to generate and exchange secrets" in *Proceedings of the 27th Annual Symposium on Foundations of Computer Science, SFCS* (IEEE Computer Society, Washington, DC, 1986), vol. 86, pp. 162–167.
5. C. Gentry, "Fully homomorphic encryption using ideal lattices" in *Proceedings of the Forty-first Annual ACM Symposium on Theory of Computing, STOC '09* (Association for Computing Machinery, New York, NY, 2009), pp. 169–178.
6. H. Cho, D. J. Wu, B. Berger, Secure genome-wide association analysis using multiparty computation. *Nat. Biotechnol.* **36**, 547–551 (2018).
7. K. Sikorska, E. Lesaffre, P. F. J. Groenen, P. H. C. Eilers, GWAS on your notebook: Fast semi-parallel linear and logistic regression for genome-wide association studies. *BMC Bioinf.* **14**, 166 (2013).
8. J. H. Cheon, A. Kim, M. Kim, Y. Song, "Homomorphic encryption for arithmetic of approximate numbers" in *Advances in Cryptology – ASIACRYPT 2017*, T. Takagi, T. Peyrin, Eds. (Springer International, Cham, Switzerland, 2017), pp. 409–437.
9. L. G. Fritsche *et al.*, A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nat. Genet.* **48**, 134–143 (2016).
10. C-Y. Chen *et al.*, Improved ancestry inference using weights from external reference panels. *Bioinformatics* **29**, 1399–1406 (2013).
11. S. Purcell, PLINK. http://zzz.bwh.harvard.edu/plink/. Accessed 18 October 2019.
12. G. Asharov *et al.*, "Multiparty computation with low communication, computation and interaction via threshold fhe" in *Advances in Cryptology – EUROCRYPT 2012,* D. Pointcheval, T. Johansson, Eds. (Springer, Berlin, Germany, 2012), pp. 483–501.
13. J. C. Denny *et al.*, Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31**, 1102–1111 (2013).
14. M. R. Nelson *et al.*, The genetics of drug efficacy: Opportunities and challenges. *Nat. Rev. Genet.* **17**, 197–206 (2016).
15. A. Almeida *et al.*, A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499–504 (2019).
16. M. Albrecht *et al.*, "Homomorphic encryption security standard" (Toronto, ON, Canada, 2018).
17. Z. Brakerski, C. Gentry, V. Vaikuntanathan, "(leveled) fully homomorphic encryption without bootstrapping" in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, ITCS '12* (Association for Computing Machinery, New York, NY, 2012), pp. 309–325.
18. J. Fan, F. Vercauteren, Somewhat practical fully homomorphic encryption. Cryptology ePrint Archive:2012/144 (17 March 2012).
19. J.-C. Bajard, J. Eynard, M. A. Hasan, V. Zucca, "A full RNS variant of FV like somewhat homomorphic encryption schemes" in Selected Areas in Cryptography – SAC 2016, R. Avanzi, H. Heys, Eds. (Springer International, Cham, Switerland, 2017), pp. 423–442.
20. S. Halevi, Y. Polyakov, V. Shoup, "An improved RNS variant of the BFV homomorphic encryption scheme" in *Topics in Cryptology – CT-RSA 2019,* M. Matsui, Ed. (Springer International, Cham, Switzerland, 2019), pp. 83–105.
21. M. Blatt, A. Gusev, Y. Polyakov, K. Rohloff, V. Vaikuntanathan, Optimized homomorphic encryption solution for secure genome-wide association studies. Cryptology ePrint Archive:2019/223 (1 April 2019).
22. J. H. Cheon, K. Han, A. Kim, M. Kim, Y. Song, "Bootstrapping for approximate homomorphic encryption" in *Advances in Cryptology – EUROCRYPT 2018*, J. B. Nielsen, V. Rijmen, Ed. (Springer International, Cham, Switzerland, 2018), pp. 360–384.
23. H. Chen, I. Chillotti, Y. Song, "Improved bootstrapping for approximate homomorphic encryption" in *Advances in Cryptology – EUROCRYPT 2019*, Y. Ishai, V. Rijmen, Eds. (Springer International, Cham, Switzerland, 2019), pp. 34–54.
24. K. Han, M. Hhan, J. H. Cheon, Improved homomorphic discrete fourier transforms and fhe bootstrapping. *IEEE Access* **7**, 57361–57370 (2019).
25. Y. Polyakov, K. Rohloff, G. W. Ryan, D. Cousins, PALISADE Lattice Cryptography Library (Release 1.7.4, 2020).
26. J. H. Cheon, A. Kim, D. Yhee, Multi-dimensional packing for heaan for approximate matrix arithmetics. Cryptology ePrint Archive:2018/1245 (21 December 2018).
27. P. D. Sasieni, From genotypes to genes: Doubling the sample size. *Biometrics* **53**, 1253–1261 (1997).
28. K. T. Zondervan, L. R. Cardon, The complex interplay among factors that influence allelic association. *Nat. Rev. Genet.* **5**, 89–100 (2004).
29. M. Blatt, A. Gusev, Y. Polyakov, S. Goldwasser, Prototypes for secure large-scale genome-wide association studies using homomorphic encryption. GitLab. https://gitlab.com/duality-technologies-public/palisade-gwas-demos/. Deposited 24 March 2020.