PREMIX: PRivacy-preserving EstiMation of Individual admiXture

Feng Chen, Ph.D.[1*], Michelle Dow[1*], Sijie Ding[2], Yao Lu[2], Xiaoqian Jiang, Ph.D.[1], Hua Tang, Ph.D.[3], Shuang Wang, Ph.D.[1]
[1]Department of Biomedical Informatics, UC San Diego, La Jolla, CA
[2]Department of Electrical and Computer Engineering, UC San Diego, La Jolla, CA
[3]Department of Genetics, Stanford University, Stanford, CA

Abstract

In this paper we proposed a framework: PRivacy-preserving EstiMation of Individual admiXture (PREMIX) using Intel software guard extensions (SGX). SGX is a suite of software and hardware architectures to enable efficient and secure computation over confidential data. PREMIX enables multiple sites to securely collaborate on estimating individual admixture within a secure enclave inside Intel SGX. We implemented a feature selection module to identify most discriminative Single Nucleotide Polymorphism (SNP) based on informativeness and an Expectation Maximization (EM)-based Maximum Likelihood estimator to identify the individual admixture. Experimental results based on both simulation and 1000 genome data demonstrated the efficiency and accuracy of the proposed framework. PREMIX ensures a high level of security as all operations on sensitive genomic data are conducted within a secure enclave using SGX.

Introduction

Identifying the demographic histories of patients is an important problem arising in biomedical research. For example, given the accurate ethnicity information, researchers can better understand whether certain populations are more susceptible to particular disease or most likely to benefit from certain therapeutic interventions[1]. Understanding the individual admixture from different ancestries is also important for researchers who conduct case-control association studies[2]. Electronic medical records (EMRs) can provide clinicians and biomedical researchers a new perspective in studying associations with the symptom or medication use. However, research studies based on the races/ethnicity from EMRs often faces problem with missing or inaccurate self-described information[3]. Hispanics, for example, represent an admixed group between Native American, Caucasian and African. In addition, African-Americans represent another large admixed group. Researchers have shown that the individuals within the Hispanics or African-Americans groups did not form a distinct subgroup, but clustered variously within the other groups[4]. As a result, the self-report ethnicity information in EMRs may not provide the most accurate characterization of patients.

Genome-wide association studies (GWAS) provide a powerful tool for identifying genetic biomarkers which reflects an individual's ethnicity by applying admixture models on allele frequencies of SNPs[5,6]. A basic assumption for ethnicity testing is that any current individual genome or population is a mixture of ancestries from past populations[7]. Population methods developed according to the amount of loci that can be traced back to a certain ancestry population is largely used. Companies such as 23andMe[8] or Ancestry DNA[9] have been the major autosomal DNA tests existed to reveal the ancestry of an individual. However, it is usually infeasible for researchers to scan for every patient's ethnicity through these expensive tests. Rapid advances in sequencing technologies enable the meaningful use of human genomic data in a wide range of healthcare and biomedicine applications. Reuse existing genomic data of patients to identify patient ethnicity or improve the accuracy of self-report information can significantly improve the data quality in research study that requires population stratification.

The research team of 23andMe published 22 population-specific common SNPs that can reflect demographic histories[5]. The study was done from the self-reporting, participant-driven data gathered on the Web, and associations were discovered for the hair color, eye color, and freckling (in the genes OCA2, HERC2, SLC45A2, SLC24A4, IRF4, TYR, TYRP1, ASIP, and MC1R)[5]. Similar researches with SNPs associations are done by Yaeger *et al.*[10] and Kosoy *et al.*[11], which found 107 and 128 SNP race/ethnicity-related biomarkers, respectively. For example, Yaeger *et al.*[10] investigated with 50 African Americans and 40 Nigerians as their subjects. Ancestry informative markers (AIMs) used in their study were based on bi-allelic SNPs that were selected from the Affymetrix 100K SNP chip based on "informativeness"[12] of ancestry's genotype data. Informativeness[12] between multiple population groups

[*] Both authors share the first authorship

was determined using mutual information. Furthermore, Kosoy *et al.*[11] worked on providing continental ancestry and characterized a set of 128 AIMs. The markers were chosen for informativeness, genome-wide distribution, and genotype reproducibility from 825 individuals. There are several ancestry estimation software that quantify genetic variation of admixture between populations using high-throughput sequencing data, such as the models used in the programs STRUCTURE[13], FRAPPE[14], TESS[15], and Admixture[16].

However, many existing studies on race/ethnicity identification are restricted by sample size or biased by sample selection. For example, the evaluations for Eriksson *et al.*[17] were done only on the European population in America. Yaeger *et al.*[10] specifically focused on African Americans born in the United States and in Africa. A total of 825 individuals were examined by Kosoy *et al.*[11] covering a wider range of individuals, but the sample size is still limited. Aggregating data from multiple sources could significantly improve the power of the study in race/ethnicity identification. However, direct sharing of labeled patients' genetic information for data mining would violate the policies concerning patient privacy[18]. Besides the privacy concerns in data mining phase, the same issues are also associated with the testing phase, where a researcher needs to identify the ethnicity through individual's genomic data, but without compromising the patient's privacy.

Regarding the privacy concern, human genomic data must be handled carefully to avoid disclosure of sensitive patient information to unauthorized parties. Previous studies[19–23] have demonstrated several privacy risks regarding to human genomic data. For example, Homer *et al.*[24] discovered that the presence of an individual in a case group can be reliably determined (known as a re-identification attack) from the allele frequencies using an individual's DNA profile, which can be acquired, for example, from a single hair or a drop of blood. The biomedical community has recognized the importance of privacy and data protection for genomic projects[25]. Many privacy and security technologies, e.g., differential privacy (DP)[26], homomorphic encryption (HME)[27–30] and secure multiparty computation (SMC)[31,32] have advanced in protecting biomedical data[33–39]. Among them, DP solutions will alter data to make it difficult to identify information to a particular individual, and DP might also render outputs useless[25]. HME and SMC (e.g. based on garbled circuits and secret sharing) hold the promise of secure general-purpose computing in the cloud but existing solutions are too computationally cumbersome to be used for complex big-data analysis. In addition, efficient SMC solutions exist (e.g. based on secret sharing and arithmetic circuit) but are domain-specific, thus inappropriate for exploratory analyses that need constant tuning. Distinct from some of the existed tools, our pipeline will not only allow users to perform ethnicity detection of a patient, but also provide secure protection of the subject's information (Figure 1). We chose to use Intel® Software Guard Extensions (Intel® SGX)[40], which is a set of new CPU instructions that can be used by applications to set aside private regions of code and data. Intel® SGX allows application to protect sensitive data from unauthorized access or modification and enables application to preserve the confidentiality without disrupting the software system. Both the labeled data and the unlabeled data from different sources will be protected during the process (Figure 1). Clients can only obtain the encrypted results (e.g., estimation of individual admixture) but cannot access the training data deposited by other institutions.

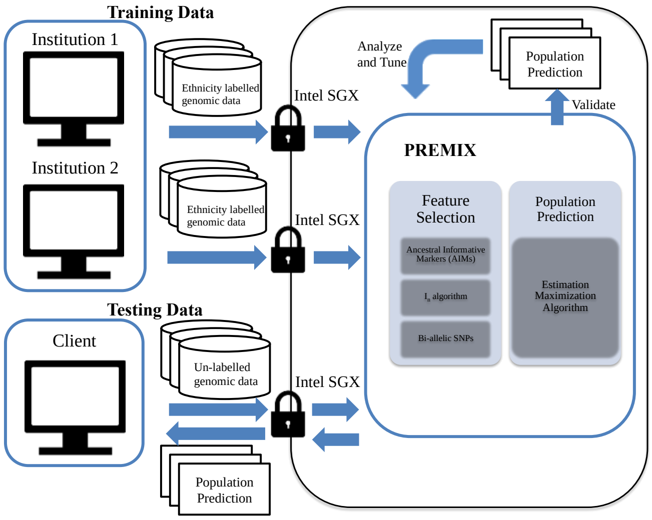


Figure 1. Workflow of the proposed PREMIX framework.

Materials and Method

In this paper, we will focus on the secure estimation of individual admixture using genomic data. We will first introduce the methods used for selecting AIMs and identifying individual admixture followed by the details about Intel® SGX frameworks.

<u>*Selection of AIMs:*</u> The selection of the most discriminative AIMs can significantly improve the efficiency of prediction for identifying individual admixture[10]. In this paper, we compute and sort the mutual information[12] between multiple population groups to select the most discriminative bi-allelic SNPs (used as AIMs in this study).

More specifically, the markers that are informative in one collection of source populations are generally informative in others[10]. Therefore, we apply statistical methods that use multilocus genotypes and population allele frequencies to represent the average frequency of allele at a certain locus, and assign informativeness to no-admixture model based on the conditional entropy of a random population given the knowledge of the genotype. Next, the informativeness is sorted from the largest to the smallest.

Identifying individual admixture: Our pipeline determines an individual's ethnicity based on the method developed by Tang *et al.*[14]. The Estimation Maximization (EM) algorithm proposed by this model demonstrates increased robustness and comparable efficiency when compared to existing maximum likelihood (ML) model[14] and Bayesian MCMC method[13]. This estimation allows for uncertainty in ancestral allele frequencies, provides an advantage toward separating an admixture population at an individual level, which they referred to as "individual admixture" (IA), and could achieve extensive stimulations to produce point estimates[14]. The goal is to estimate IA for the admixed individuals, $Q_i = (q_{i1}, \ldots, q_{iK})$ and $i = 1, \ldots I_0$, where Q_i is the individual admixture fraction of the i-th individual; I_0 is the total number of individuals to be identified; K is the number of ancestries. The allele frequencies of different ancestral markers are denoted as P. A likelihood function given an unobservable variable $Z_{ima} \in \{1, \ldots, K\}$ can be expressed as

$$l(G, Z|P, Q) = \sum_{i=1}^{I} \sum_{m=1}^{M} \sum_{a=1}^{2} \sum_{l=1}^{L_m} \sum_{k=1}^{K} \mathbf{1}(G_{ima} = l, Z_{ima} = k) \log(p_{mlk} q_{ik})$$

where $\mathbf{1}(\cdot)$ is an indicator function; p_{mlk} is the frequency of allele l at marker m for the k-th ancestry; G_{ima} is the allele for the i-th individual at the a-th allele of marker m. An EM algorithm can be implemented to efficiently estimate the parameters P and Q. Convergence is declared when the difference in the estimates of Q and P fall below a small threshold[14].

Secure computation methods: To mitigate the privacy risk while supporting scientific discovery, security researchers have developed many theoretical frameworks. However, even the best-known methods based on HElib[41] for homomorphic encryption or FlexSC[42] for garbled circuit-based secure multiparty computation are not practical enough to handle large-scale genomic data analysis. These real challenges motivate the development of new solutions. Intel recently announced a new Software Guard eXtensions (SGX)[40] architecture for their next generation CPUs, which shed light to novel solutions to above mentioned challenges using a hybrid software-hardware framework. Figure 2 shows some of the major conceptual functions of the SGX architecture (need to be developed on a case-by-case manner depending on the application). As a built-in feature in the Intel® Skylake family Central Processing Units (CPUs), SGX-enabled devices can be found in most recently released computing platforms (e.g., laptops, desktops and servers). Intel® SGX framework provides a cost-effective solution to achieve affordable secure computation in terms of both complexity and finance concerns. For example, a pioneer study of SGX-based MapReduce framework for distributed high-performance computations[43] shows a negligible overhead of 8% to achieve read/write integrity using SGX. This is a significant advantage of using SGX in comparison to other secure computation scheme like homomorphic encryption or garbled circuits, which usually increase the complexity thousands of times over. Furthermore, an SGX-enabled machine only cost as low as a few hundred dollars. In SGX, a protected area in CPU, which is usually referred to as enclave, is dedicated to execute sensitive codes and compute sensitive data in a secure manner, where any interfere from software outside the enclave are prohibited by the SGX hardware. Therefore, both data confidentiality and integrity can be achieved with a proper systemic design of SGX applications. SGX is resilient to both software level attacks (e.g., malicious operating system, etc.) and hardware level attacks (e.g., for memory, hard disk, network etc.). Some preliminary studies[43–45] have revealed the possibility of SGX to significantly enhance the security and privacy of many applications. However, most of them are based on simulation study and none of them has tackled genomic data security and privacy in a real SGX-enabled computing platform.

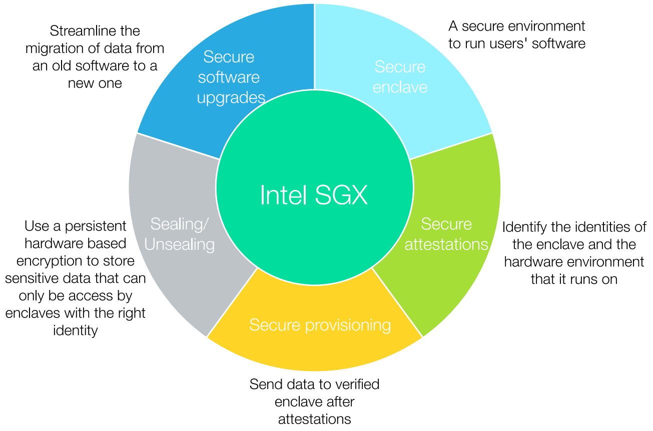

Streamline the migration of data from an old software to a new one
Secure software upgrades
Secure enclave
A secure environment to run users' software
Intel SGX
Sealing/ Unsealing
Use a persistent hardware based encryption to store sensitive data that can only be access by enclaves with the right identity
Secure attestations
Identify the identities of the enclave and the hardware environment that it runs on
Secure provisioning
Send data to verified enclave after attestations

Figure 2. Conceptual functions of the SGX architecture.

To ensure the security of the whole system, an SGX framework requires the adoption of industry-standard cryptographic primitives and systemic implementation of several key steps, as shown in Tables 1 and 2, respectively. Taking advantage of this novel architecture, we developed a PREMIX framework for privacy-preserving estimation of individual admixture in this paper.

Table 1. Summary of cryptographic primitives to be adopted in the design of SGX application.

Cryptographic primitives	**Description**	**Security**	**Industry standard**
Advanced Encryption Standard (AES) in Galois Counter Mode (GCM)	Authenticated encryption, which provides simultaneous protection of data confidentiality and authenticity.	128 bits	NIST SP 800-38D guideline [46]
Elliptic Curve Diffie–Hellman (ECDH)	A key agreement protocol based on Elliptic Curve Cryptography (ECC) to establish securely shared symmetric key for AES over on an insecure channel.	256 bits	NIST SP 800-56A guideline [47]
Elliptic Curve Digital Signature Algorithm (ECDSA)	An ECC based digital signature scheme to ensure the source of data is as claimed.	256 bits	FIPS Pub. 186-3 guideline[48]

Table 2. Key steps and their corresponding cryptographic primitives for SGX application to achieve efficient and trustworthy computation.

Key steps	**Description**	**Cryptographic primitives**		
		AES-GCM	**ECDH**	**ECDSA**
Remote attestation	Securely provision an enclave from an authorized user who is outside the computing platform		X	X
Data provision	Securely transfer sensitive data into enclave	X	X	

Results

Study Subjects and Ancestral Population: We have used the phase 3 release (May 2013) of the 1000 Genomes project data to identify SNPs. Datasets from 1000 Genome project include genetic variation across diverse populations from Europe, Asia, Africa and the Americas. The present 1000 Genome data contain 2504 samples from 26 populations which can be categorized into five super-populations: East Asian (EAS), South Asian (SAS), African (AFR), European (EUR), and American (AMR). The global allele frequencies for each super-populations were calculated by counting the AC ("Total number of alternate alleles in called genotypes") and AN ("Total number of alleles in called genotypes") for all the individuals from a particular super population and using that to calculate the allele frequencies.

Experimental setup: We use both simulation and real data to test the performance of our algorithm. For simulation experiment, we follow the setup used by Tang et al.[14]. Assume there are two ancestral populations X and Y. The simulated data consist of 500 admixed individuals, and 250 individuals from each of two populations as the training data. Since there are two groups, the IA vector is a scalar, and we sample it from a mixture model of Gaussian and uniform distributions. We select the SNPs that the differences of its allele frequencies from two parties are bigger than a δ-value 0.3. Conditioning on the SNPs and the IA vector, we can randomly generate the 1000 individuals for simulation. From 1000 genome data, we used two populations: ACB (African Caribbeans in Barbados) and TSI (Toscani in Italia), and each of them contains 96 and 107 subjects. We extracted the first 31,000 SNPs of their 22nd chromosomes. All of these SNPs will be processed by the informativeness algorithm[12] to choose the top SNPs for computing the IA vector of the admixed individuals.

We implemented the proposed PREMIX server and client on two machines: the server is an Intel® Xeon core E3-1275 v5 with Intel® SGX support and 64 GB memory; the client machine is Intel® Core i7-6820HQ CPU and 48GB Memory.

The experiments are designed to focus on the following three aspects: (a) computational complexity comparison between secure SGX-based C++ implementation vs insecure C++ implementation; (b) the simulation data results; (c) the real data results using different number of top informative AIMs.

Experimental results: Table 3 shows the key steps and total running times of PREMIX using secure SGX implementations with encrypted remote data and insecure C++ with local data. We tested four different data sizes, and all of the results in Table 3 are the averaging values over 10 trials. From the results, we can see that, there is no significant different between the two frameworks in implementing EM algorithm, but the total running times of secure SGX is a little slower than those of insecure C++. The additional overhead in the total running times of secure SGX is due to the data encryption, attestation, data transfer and analysis over encrypted data.

Table 3. Comparison of the running times of computing PREMIX between secure SGX-based C++ implementation vs. insecure C++ implementation. Here I_0 and I are the number of admixed individuals and the total individuals, respectively. The unit of all running times is second.

I_0/I	Secure SGX-based C++					Insecure C++	
	Client data encryption	**Attestation**	**Data transfer**	**EM algorithm**	**Total**	**EM algorithm**	**Total**
250/500	<0.001	0.563	0.453	1.827	3.788	2.186	3.296
500/1000	0.006	0.567	0.766	3.613	6.631	4.042	5.847
750/1500	0.008	0.579	1.040	6.937	10.987	6.873	9.422
1000/2000	0.016	0.569	1.449	9.424	14.662	9.095	12.539

Simulated data is used for evaluating the PREMIX framework. We suppose that the labeled data with ethnicity information are non-admixed and there are two populations. Each individual in the labeled data set is sampled from one of the two populations. The number of SNPs used in our simulation is 200. The iteration rounds of the EM algorithm are set to 200.

Table 4. Performance based on simulated data.

δ	RMSE	Bias
0.3	0.117	0.015
0.4	0.089	0.006
0.5	0.080	0.010
0.6	0.070	0.001

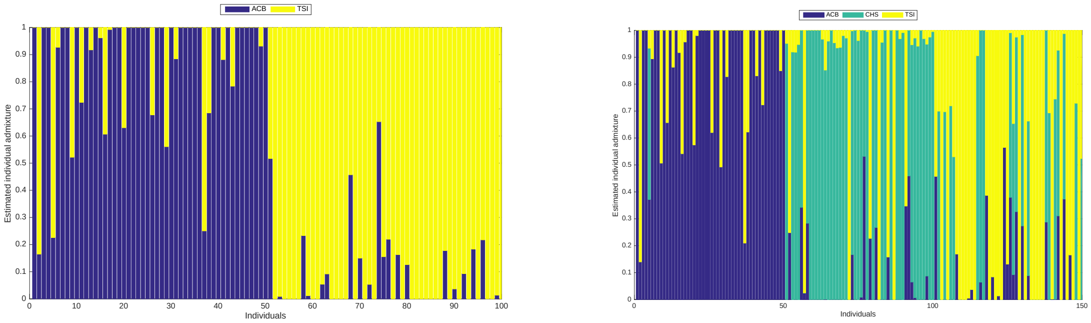
Table 5. Percentage of correctly identified individuals using different number of AIMs.

M	Percentage of correct identification
10	81 %
20	93 %
50	95 %
100	95 %
200	95 %
500	97 %

Table 6. Percentage of correctly identified individuals using different labeled data sizes.

Labeled data size	Percentage of correct identification
20	93%
40	95%
60	95%
80	95%
100	95%

Table 4 is the results of PREMIX using simulated data. Since we know the ground truth under simulation environment. We can compute the root mean square error (RMSE) and the bias. Table 4 shows that the RMSE will less than 0.1 with no more than 0.01 bias, if the δ value of two groups is bigger than 0.3.



(a) Two populations

(b) Three populations

Figure 3. Estimated individual admixture using the proposed PREMIX framework.

For the real data, we split each population into two groups, where the first group is used as admixed individuals to be predicted and the second group is used as the labeled data. Specifically, the individuals from ACB and TSI populations were included in our experiment. For both populations, the first 50 individuals are viewed as admixed individual, and the rest individuals from both populations (i.e., 46 ACB individuals and 57 TSI individuals) are used as labeled data. In addition, we suppose these labeled data were from two sources to simulate a secure collaboration scenario. The SNPs were screened based on their informativeness[12], where the top M SNPs will be selected for the next step in PREMIX framework. For the real data, since there is no ground true of IA vector, the admixed individual will be classified to the population based on their maximum estimated IA component. Tables 5 and 6 depict the percentage of correctly identified individuals using different number of AIMs and different labeled data sizes, respectively. In Table 5, we can see that the percentage of correct identification increases as the number of AIMs M increases. Based on our experiments, a high percentage of correct identification can be achieved with 50 or more AIMs. Moreover, Table 6 shows the framework can achieve a relative accurate identification performance with as few as 40 labeled data.

Figure 3 (a) shows the estimated IA vectors for two populations. The first 50 individuals are from the ACB population, and the second 50 individuals are from TSI population. We can see that the PREMIX can successfully identify the ethnicities of most individuals in both groups.

To further evaluate the performance of the PREXIM, we included a third population CHS (Southern Han Chinese) in our experiment. In Figure 3 (b), the first 50 individuals are from the ACB; the second 50 individuals are from CHS; and the final 50 individuals are from TSI. We see that there is some performance degradation of the proposed PREXIM framework in identifying more than two populations.

Discussion and Limitation

The main contribution of this paper is to introduce a new hybrid solution (i.e., Intel® SGX) using both hardware and software to enable efficient and privacy-preserving estimation of individual admixture. The proposed PREMIX framework can protect the privacy of sensitive genomic data with ancestry information, as well as the privacy of the data users, who would like to identify their individual admixture. Due to the adoption of strong security protection primitives, multiple data owners can collaborate on the study to improve the estimation performance without sacrificing individual data privacy. In the proposed framework, we provided both a secure feature selection module based on informativeness of SNPs and a secure EM based maximum likelihood estimator to achieve both computational efficiency and estimation accuracy. Our experimental results demonstrated the advantage of secure collaboration in identifying individual admixture.

There are several limitations in this study. First, even it can well protect the data privacy, the SGX hardware is vulnerable to Denial-of-service (DoS) attack; however, the data privacy will not be compromised under this attack. Second, proposed method was only evaluated through limited data sets (i.e., simulated data and 1000 genome data) in this pilot study. The use of Human Genome Diversity project (HGDP) could improve the impact of this study and provide better performance assessment. In addition, the proposed method relies on an EM-based maximum likelihood estimator, which can only support a small number of populations. Recently, many advanced ethnicity identification programs have been developed particularly for genome-wide SNP data. For example, TESS3[63] is an updated version of the spatial ancestry estimation program TESS, which combines matrix factorization and spatial statistical methods. TESS3 estimates ancestry coefficients with comparable accuracy and fast run-times, and can be used to perform genome scans for selection, separate adaptive from non-adaptive genetic variation using ancestral allele frequency differentiation tests. AncestryMapper[50] assigns each individual analyzed with a genetic identifier, referred as Ancestry Mapper Id (AMid) which corresponds to its relationship to the HGDP reference population. TreeMix[51] follows a tree-based approach with branches built by maximum likelihood lengths and migration weights. Population was identified by searching through the space of possible graphs with optimized the branch lengths and weights. FastSTRUCTURE[52] is a recent modification of the popular model STRUCTURE which provides a faster approximate inference using a variational Bayesian framework and poses the problem of computing relevant posterior distributions as an optimization problem. The software identifies the number of populations represented in a dataset with heuristic and new hierarchical prior to detect weak population structure in the data. Among a large population, algorithm such as Eagle[53] detects association analysis of rare variants from a large population cohorts based on genotyping arrays using long-range phasing (LRP) to rapidly phase segments of genome identical-by-descent (IBD) with closely or distantly related individuals. Eagle runs two iterations of fast approximate Viterbi decoding using a simple diploid analog of the Li-Stephens HMM to allows phasing of segments lacking IBD to ensure accurate results. The development of trustworthy computation framework to support advanced methods in race/ethnicity identification warrants the further investigation along this line. Finally, the limited secure memory (~

96 MB) in SGX restricts the algorithm to process a huge amount of data concurrently. In the next step, we will optimize the secure memory usage to improve the data processing capacity of the proposed method.

Acknowledgment

F.C., M.D., Y.L. and S.W. contributed to the majority of the writing. F.C. and S.W. designed the method and devised the simulation experiments. F.C., Y.L., S.D. and S.W. implemented algorithms. X.J. provided detailed edits. S.W., X.J. and H.T. provide critical suggestions to improve the paper. This work has been supported by the NHGRI (R00HG008175, R01HG007078), NIGMS (R01GM114612), NLM (R00LM011392, R21LM012060), NHLBI (U54HL108460), NIH bioCADDIE (1U24AI117966-01).

References

1. Rowell JL, Dowling NF, Yu W, Yesupriya A, Zhang L, Gwinn M. Trends in population-based studies of human genetics in infectious diseases. PLoS One. 2012;7(2):e25431.
2. Tang H, Quertermous T, Rodriguez B, Kardia SLR, Zhu X, Brown A, et al. Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies. Am J Hum Genet. 2005;76(2):268–75.
3. Smith N, Iyer RL, Langer-Gould A, Getahun DT, Strickland D, Jacobsen SJ, et al. Health plan administrative records versus birth certificate records: quality of race and ethnicity information in children. BMC Health Serv Res. 2010;10(1):1.
4. Risch N, Burchard E, Ziv E, Tang H. Categorization of humans in biomedical research: genes, race and disease. Genome Biol. 2002;3(7):1–12.
5. Choudhury A, Hazelhurst S, Meintjes A, Achinike-Oduaran O, Aron S, Gamieldien J, et al. Population-specific common SNPs reflect demographic histories and highlight regions of genomic plasticity with functional relevance. BMC Genomics. 2014;15(1):1.
6. Pfaff CL, Barnholtz-Sloan J, Wagner JK, Long JC. Information on ancestry from genetic markers. Genet Epidemiol. 2004;26(4):305–15.
7. Pugach I, Matveyev R, Wollstein A, Kayser M, Stoneking M, others. Dating the age of admixture via wavelet transform analysis of genome-wide data. Genome Biol. 2011;12(2):R19.
8. 23andMe [Internet]. [cited 2016 Jul 3]. Available from: https://www.23andme.com/
9. Ancestry DNA [Internet]. [cited 2016 Jul 3]. Available from: http://dna.ancestry.com/
10. Yaeger R, Avila-Bront A, Abdul K, Nolan PC, Grann VR, Birchette MG, et al. Comparing genetic ancestry and self-described race in african americans born in the United States and in Africa. Cancer Epidemiol Biomarkers Prev. 2008;17(6):1329–38.
11. Kosoy R, Nassir R, Tian C, White PA, Butler LM, Silva G, et al. Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. Hum Mutat. 2009;30(1):69–78.
12. Rosenberg NA, Li LM, Ward R, Pritchard JK. Informativeness of genetic markers for inference of ancestry. Am J Hum Genet. 2003;73(6):1402–22.
13. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000 Jun;155(2):945–59.
14. Tang H, Peng J, Wang P, Risch NJ. Estimation of individual admixture: analytical and study design considerations. Genet Epidemiol. 2005;28(4):289–301.
15. François O, Durand E. Spatially explicit Bayesian clustering models in population genetics. Mol Ecol Resour. 2010 Sep;10(5):773–84.
16. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009 Sep;19(9):1655–64.
17. Eriksson N, Macpherson JM, Tung JY, Hon LS, Naughton B, Saxonov S, et al. Web-based, participant-driven studies yield novel genetic associations for common traits. PLoS Genet. 2010;6(6):e1000993.
18. NIH Genomic Data Sharing Policy [Internet]. 2014 [cited 2015 Jun 5]. Available from: http://gds.nih.gov/03policy2.html
19. Naveed M, Ayday E, Clayton EW, Fellay J, Gunter CA, Hubaux J-P, et al. Privacy and Security in the Genomic Era. ACM Comput Surv [Internet]. 2015 May 8 [cited 2014 Aug 11];48(1):6. Available from: http://arxiv.org/abs/1405.1891
20. Loukides G, Gkoulalas-Divanis A, Malin B. Anonymization of electronic medical records for validating genome-wide association studies. Proc Natl Acad Sci U S A [Internet]. 2010/04/14 ed. 2010;107(17):7898–903. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20385806
21. Wang R, Li YF, Wang X, Tang H, Zhou X. Learning your identity and disease from research papers. In:

Proceedings of the 16th ACM conference on Computer and communications security - CCS '09 [Internet]. New York, New York, USA: ACM Press; 2009 [cited 2014 Aug 13]. p. 534–44. Available from: http://dl.acm.org/citation.cfm?id=1653662.1653726

22. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. Science (80-). 2013 Jan 18;339(6117):321–4.

23. Lin Z, Owen AB, Altman RB. Genomic research and human subject privacy. Science (80-) [Internet]. 2004 Jul 9 [cited 2014 Aug 13];305(5681):183. Available from: http://www.sciencemag.org/content/305/5681/183.short

24. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet [Internet]. 2008 Aug [cited 2015 Sep 29];4(8):e1000167. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2516199&tool=pmcentrez&rendertype=abstract

25. Jiang X, Zhao Y, Wang X, Malin B, Wang S, Ohno-Machado L, et al. A community assessment of privacy preserving techniques for human genomes. BMC Med Inform Decis Mak [Internet]. 2014 Dec 8 [cited 2015 Jan 23];14 Suppl 1(Suppl 1):S1. Available from: http://www.biomedcentral.com/1472-6947/14/S1/S1

26. Dwork C. Differential privacy. Int Colloq Autom Lang Program. 2006;4052(d):1–12.

27. Naehrig M, Lauter K, Vaikuntanathan V. Can homomorphic encryption be practical? In: Proceedings of the 3rd ACM workshop on Cloud computing security workshop - CCSW '11 [Internet]. New York, NY, USA: ACM Press; 2011 [cited 2014 Apr 30]. p. 113. Available from: http://dl.acm.org/citation.cfm?id=2046682

28. Bos JW, Lauter K, Naehrig M. Private predictive analysis on encrypted medical data. J Biomed Inform. 2014;50:234–43.

29. Lauter K, Adriana L, Naehrig M. Private Computation on Encrypted Genomic Data. :1–21.

30. Graepel T, Lauter K, Naehrig M. ML confidential: Machine learning on encrypted data. In: Information Security and Cryptology--ICISC 2012. Springer; 2013. p. 1–21.

31. Verle D Du, Kawasaki S, Yamada Y, Sakuma J, Tsuda K. Privacy-Preserving Statistical Analysis by Exact Logistic Regression. In: 2nd International Workshop on Genome Privacy and Security (GenoPri'15). San Jose, CA; 2015.

32. Kamm L, Bogdanov D, Laur S, Vilo J. A new way to protect privacy in large-scale genome-wide association studies. Bioinformatics [Internet]. 2013 Apr 1 [cited 2014 Aug 11];29(7):886–93. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3605601&tool=pmcentrez&rendertype=abstract

33. Naveed M, Ayday E, Clayton EW, Fellay J, Gunter CA, Hubaux J-P, et al. Privacy and Security in the Genomic Era. ACM Comput Surv [Internet]. 2015 May 8 [cited 2014 Aug 11];48(1):6. Available from: http://arxiv.org/abs/1405.1891

34. Chen F, Mohammed N, Wang S, He W, Cheng S, Jiang X. Cloud-Assisted Distributed Private Data Sharing. In: The ACM Conference on Bioinformatics, Computational Biology, and Health Informatics. Atlanta, GA; 2015.

35. Ayday E, Raisaro JL, Hubaux J-P. Personal use of the genomic data: privacy vs. storage cost. In: IEEE Global Communications Conference, Exhibition and Industry Forum--GLOBECOM. 2013.

36. Xie W, Kantarcioglu M, Bush WS, Crawford D, Denny JC, Heatherly R, et al. SecureMA: protecting participant privacy in genetic association meta-analysis. Bioinformatics. 2014;31(23).

37. Chen F, Wang S, Mohammed N, Cheng S, Jiang X. PRECISE:PRivacy-prEserving Cloud-assisted quality Improvement Service in hEalthcare. IEEE Int Conf Syst Biol [proceedings] IEEE Int Conf Syst Biol [Internet]. 2014 Oct [cited 2016 Jul 3];2014:176–83. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26146645

38. Wang S, Mohammed N, Chen R. Differentially private genome data dissemination through top-down specialization. BMC Med Inform Decis Mak [Internet]. 2014 Dec 8 [cited 2015 Mar 30];14(Suppl 1):S2. Available from: http://www.biomedcentral.com/1472-6947/14/S1/S2

39. Yu F, Ji Z. Scalable Privacy-Preserving Data Sharing Methodology for Genome-Wide Association Studies: An Application to iDASH Healthcare Privacy Protection Challenge. BMC Med Inform Decis Mak. 2014;14(Suppl 1):S3.

40. Intel® Software Guard Extensions (Intel® SGX) [Internet]. Available from: https://software.intel.com/en-us/isa-extensions/intel-sgx

41. S. Halevi, Shoup V. HElib [Internet]. [cited 2015 Aug 5]. Available from: https://github.com/shaih/HElib

42. Wang X. A Flexible Efficient Secure Computation Backend computer program (College Park, Department of Computer Science University of Maryland) [Internet]. 2015 [cited 2016 Feb 1]. Available from:

https://github.com/yhuang912/FlexSC

43. Schuster F, Costa M, Fournet C, Gkantsidis C, Peinado M, Mainar-Ruiz G, et al. VC3: Trustworthy data analytics in the cloud using SGX. In: Security and Privacy (SP), 2015 IEEE Symposium on. 2015. p. 38–54.
44. Kim S, Shin Y, Ha J, Kim T, Han D. A First Step Towards Leveraging Commodity Trusted Execution Environments for Network Applications. In: Proceedings of the 14th ACM Workshop on Hot Topics in Networks. 2015. p. 7.
45. Baumann A, Peinado M, Hunt G. Shielding applications from an untrusted cloud with haven. In: USENIX Symposium on Operating Systems Design and Implementation (OSDI). 2014.
46. Dworkin MJ. SP 800-38D. Recommendation for block cipher modes of operation: Galois/Counter Mode (GCM) and GMAC. National Institute of Standards & Technology; 2007.
47. Barker E, Johnson D, Smid M. NIST special publication 800-56A: Recommendation for pair-wise key establishment schemes using discrete logarithm cryptography (revised). Comput Secur Natl Inst Stand Technol (NIST), Publ by NIST. 2007;
48. Locke G, Gallagher P. FIPS PUB 186-3: Digital Signature Standard (DSS). Fed Inf Process Stand Publ. 2009;
49. Caye K, Deist TM, Martins H, Michel O, François O. TESS3: Fast inference of spatial population structure and genome scans for selection. Molecular Ecology Resources. 2015 Sep;
50. Magalhães TR, Casey JP, Conroy J, Regan R, Fitzpatrick DJ, Shah N, et al. HGDP and HapMap analysis by Ancestry Mapper reveals local and global population relationships. PLoS One. 2012 Jan;7(11):e49438.
51. Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 2012 Jan;8(11):e1002967.
52. Raj A, Stephens M, Pritchard JK. fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics. 2014 Jun;197(2):573–89.
53. Loh P-R, Palamara PF, Price AL. Fast and accurate long-range phasing and imputation in a UK Biobank cohort. bioRxiv. Cold Spring Harbor Labs Journals; 2015 Oct.