PROGRESS OF HOMOMORPHIC
ENCRIPTION FOR PROTECTING GENOMIC
DATA PRIVACY AND SECURITY IN THE PAST
4 YEARS IDASH COMPETITION

SHUANG WANG, PHD
DEPARTMENT OF BIOMEDICAL INFORMATICS
UNIVERSITY OF CALIFORNIA SAN DIEGO

SUPPORTED BY
NHGRI R00HG008175, R13HG009072
NIBIB U01EB023685

Joint work with Dr. Xiaoqian Jiang (UCSD), Dr. Lucila, Ohno-Machado (UCSD), Xiaofeng Wang (IU), Haixu Tang (IU)
• Genome data have been widely used in biomedical research
• But genomic data are also highly sensitive
  • Diseases association: predisposition to Diabetes, Cancer...
  • Re-identification: name...
  • Information disclosure of blood relatives
  • A great fear of unknown
<table>
<thead>
<tr>
<th>Author</th>
<th>Year</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sweeney [6]</td>
<td>2000</td>
<td>Identifying 87% of US citizens with the combination of ‘ZIP code, gender, date of birth’</td>
</tr>
<tr>
<td>Gottlib [7]</td>
<td>2001</td>
<td>Finding employees who are susceptible to genetic diseases depending on genomic data</td>
</tr>
<tr>
<td>Lin et al. [8]</td>
<td>2004</td>
<td>Identifying a person by as few as 75 independent SNPs</td>
</tr>
<tr>
<td>Homer et al. [9]</td>
<td>2008</td>
<td>Detecting if an individual is present in a DNA mixture within a case group</td>
</tr>
<tr>
<td>Sankararaman et al. [10]</td>
<td>2009</td>
<td></td>
</tr>
<tr>
<td>Wang et al. [11]</td>
<td>2009</td>
<td>Re-identifying individuals and reconstructing allele frequencies from research papers</td>
</tr>
<tr>
<td>Gymrek et al. [12]</td>
<td>2013</td>
<td>Identifying surnames by profiling short tandem repeats on the Y-chromosome</td>
</tr>
<tr>
<td>Claes et al. [13]</td>
<td>2014</td>
<td>Reconstructing a 3D face from human genomic data</td>
</tr>
<tr>
<td>Shringapure et al. [14]</td>
<td>2015</td>
<td>Identifying participants from using Beacon services with limited number of queries</td>
</tr>
<tr>
<td>Harmanci et al. [15]</td>
<td>2016</td>
<td>Linking phenotype and genotype data to reveal private information</td>
</tr>
<tr>
<td>Lippert et al. [16]</td>
<td>2017</td>
<td>Identification of individuals by trait prediction using whole-genome sequencing data</td>
</tr>
</tbody>
</table>
And finally, we're going to make sure that protecting patient privacy is built into our efforts from day one.
HIPAA REGULATES MEDICAL DATA SHARING

HIPAA:
Health Insurance Portability and Accountability Act

“A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable”

This method is seldom used in practice.
PROBLEMS WITH THE SAFE HARBOR METHOD

- Biometrics are Protected Health Information (PHI)
- PHI requires HIPAA

• Biometrics require HIPAA
PROBLEMS WITH THE SAFE HARBOR METHOD

- Biometrics require HIPAA
- PHI requires HIPAA
- Biometrics are Protected Health Information (PHI)
2. Under the Genomic Data Sharing (GDS) Policy, is NIH allowing investigators who are approved to download human datasets from NIH controlled-access repositories to use cloud computing?

In April 2015, NIH released the NIH Position Statement on Use of Cloud Computing Services for Storage and Analysis of Controlled-Access Data Subject to the NIH Genomic Data Sharing Policy and is now allowing investigators to request permission to transfer controlled-access genomic and associated phenotypic data obtained from NIH designated repositories under the auspices of the GDS Policy to public or private closed systems for data storage and analysis. NIH expects cloud computing systems to meet the data use and security standards outlined in NIH Security Best Practices for Controlled-Access Data Subject to the NIH Genomic Data Sharing (GDS) Policy as well as the institution’s own IT security requirements and policies. Investigators who wish to use cloud computing for storage and analysis will need to indicate in their Data Access Request (DAR) that they are requesting permission to use cloud computing, identify the cloud service provider or providers that will be employed, and describe how the cloud computing service will be used to carry out their proposed research.

The NIH strongly recommends that investigators consult with institutional IT leaders, including the Chief Information Officer (CIO) and the institutional Information Systems Security Officer (ISSO) or equivalents to develop the formal information security plan prior to receipt of controlled access data from the NIH, and institutional signing officials should validate that an appropriate security plan is in place prior to accepting liability for data loss or breach on behalf of the institution. This document provides an overview of security principles for data, access, and physical security to ensure confidentiality, privacy, and accessibility of data. This is a minimum set of requirements; additional restrictions may be needed by your institution and should be guided by the knowledge of the user community at your institution as well as your institution’s IT requirements and policies.
Data on 150,000 patients exposed in another misconfigured AWS bucket

Patient Home Monitoring failed to lock down public access to its online server, exposing personal data of 150,000 patients.

By Jessica Davis | October 12, 2017 | 02:02 PM

Kromtech Security researchers have discovered yet another unsecured Amazon S3 bucket. This time, the cloud server in question was linked to HIPAA-covered entity, Patient Home Monitoring, a vendor that provides U.S. patients with disease management services and in-home monitoring.
A COMMUNITY EFFORT OF GENOMIC DATA PRIVACY PROTECTION

2014 – 2017 iDASH genomic data privacy and security protection competition http://www.humangenomeprivacy.org
IDASH PRIVACY WORKSHOPS*

http://www.humangenomeprivacy.org/

An interdisciplinary challenge on genomic privacy research

• Motivated by real world biomedical applications and with participation of privacy technology experts, Biomedical researchers, ELSI researchers (academia and industry)

• Developed practical yet rigorous solutions for privacy preserving genomic data sharing and analysis

• Demonstrated feasibility of secure genome data analysis and dissemination using differential privacy, MPC, HE, SGX

• Reported in the media (e.g., Nature News)

* Supported by U54HL108460 initially, and then by R13HG009072

http://www.nature.com/news/extreme-cryptography-paves-way-to-personalized-medicine-1.17174
2014
- 2 countries
- 9 states
- 33 registrations
  - Privacy preserving data sharing
  - Secure release of genome analysis results

2015
- 5 countries
- 7 states
- 50+ registrations
  - Homomorphic Encryption for GWAS (MAF&Chi-Squared)
  - Secure Collaboration on DNA Analysis

2016
- 13 countries
- 10+ states
- 75+ registrations
  - Privacy-Preserving Search of Similar Cancer Patients across Organizations
  - Testing for Genetic Diseases on homomorphically Encrypted Genomes
  - Protecting queries in Beacon service

2017
- 19 countries
- 65+ Teams
  - Secure Record De-duplication
  - Secure GWAS using SGX
  - Homomorphic logistic regression
## APPLICATIONS ENABLED BY HOMOMORPHIC ENCRYPTION

<table>
<thead>
<tr>
<th>Year</th>
<th>Homomorphic encryption Applications</th>
<th>Winning Team</th>
<th>Problem setup</th>
<th>Run time</th>
<th>Peak memory cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015</td>
<td>Minor Allele Frequency</td>
<td>Stanford/MIT</td>
<td>610 SNPs and 200 individuals</td>
<td>1.847 (seconds)</td>
<td>13 (MB)</td>
</tr>
<tr>
<td></td>
<td>Chi-squared statistics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hamming Distance</td>
<td>IBM</td>
<td>100K sequences</td>
<td>472.2 (seconds)</td>
<td>2.168 (GB)</td>
</tr>
<tr>
<td></td>
<td>Approximate Edit Distance</td>
<td>Microsoft Research</td>
<td>10K sequences</td>
<td>181.92 (Seconds)</td>
<td>1.295 (GB)</td>
</tr>
<tr>
<td>2016</td>
<td>Genetic testing</td>
<td>Microsoft Research</td>
<td>(1 query (1 variant) / 50 VCF files [100k])</td>
<td>59.58 (seconds)</td>
<td>83.6 (MB)</td>
</tr>
<tr>
<td>2017</td>
<td>Logistic Regression</td>
<td>Seoul National University</td>
<td>Datasets with 1422 records and 18 features</td>
<td>10.360 (minutes)</td>
<td>2775.333 (MB)</td>
</tr>
</tbody>
</table>
2018 IDASH COMPETITION

- New challenges will be announced soon in earlier April.
- Http://www.humangenomeprivacy.org