iDASH integrating Data for Analysis, Anonymization, and SHaring

The 2nd Competition on
Critical Assessment of Data Privacy and Protection

Secure Genome Analysis

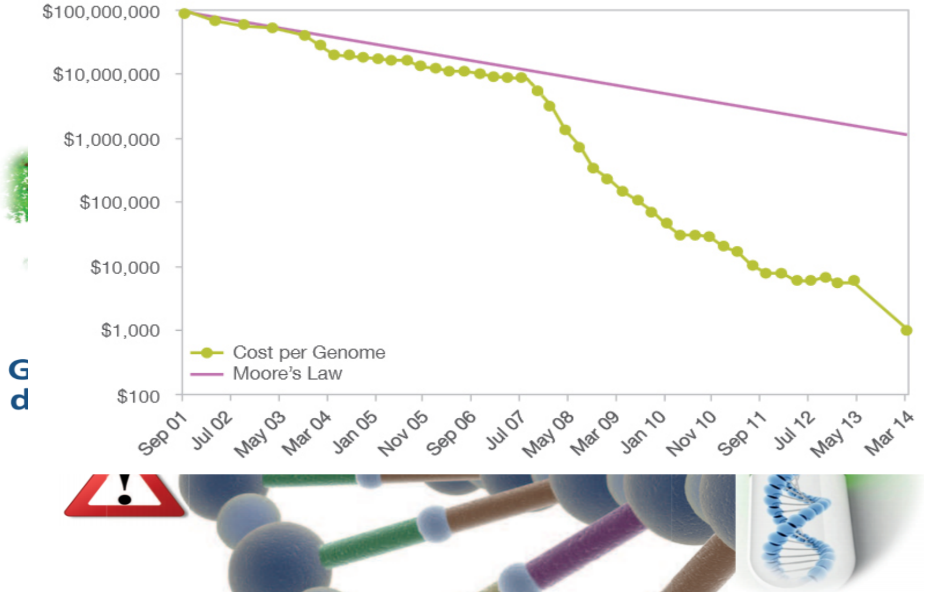
The privacy workshop is jointly sponsored by iDASH (U54HL108460) and
the collaborating R01 (R01HG007078)

Genomic Revolution

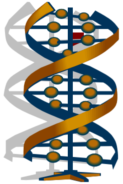
- Fast drop in the cost of genome-sequencing
  - 2000: $3 billion
  - Mar. 2014: $1,000
  - Genotyping 1M variations: below $200

- Unleashing the potential of the technology
  - Healthcare: e.g., disease risk detection, personalized medicine
  - Biomedical research: e.g., geno-phono association
  - Legal and forensic
  - DTC: e.g., ancestry test, paternity test

  ……



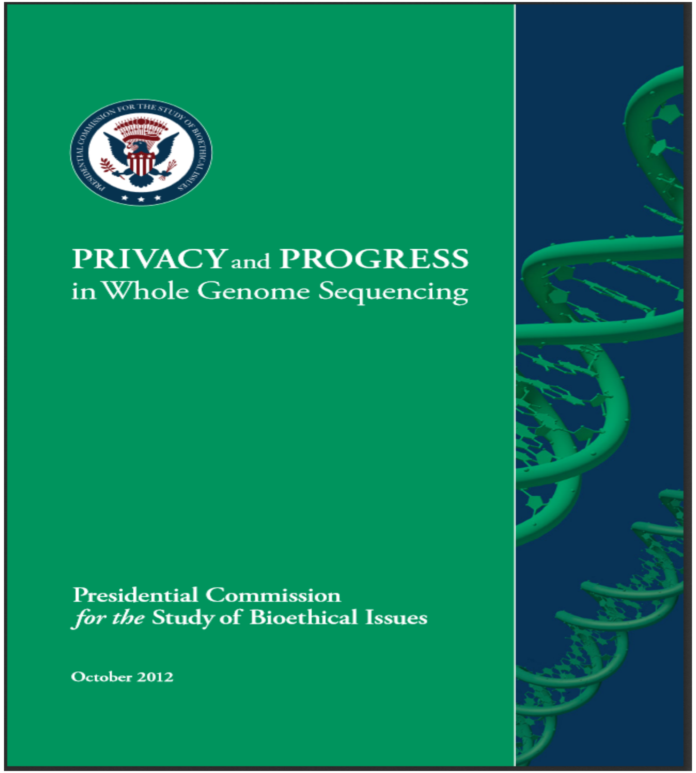
Genome Privacy



- Privacy risks
  - Genetic disease disclosure
  - Collateral damage
  - Genetic discrimination

……

- Protection
  - Clear access policies
  - Accountability
  - Data anonymization
  - Best practice for data privacy
  - Privacy awareness ……



PRIVACY and PROGRESS
in Whole Genome Sequencing

Presidential Commission
for the Study of Bioethical Issues

October 2012

For More Information

Privacy and Security in the Genomic Era

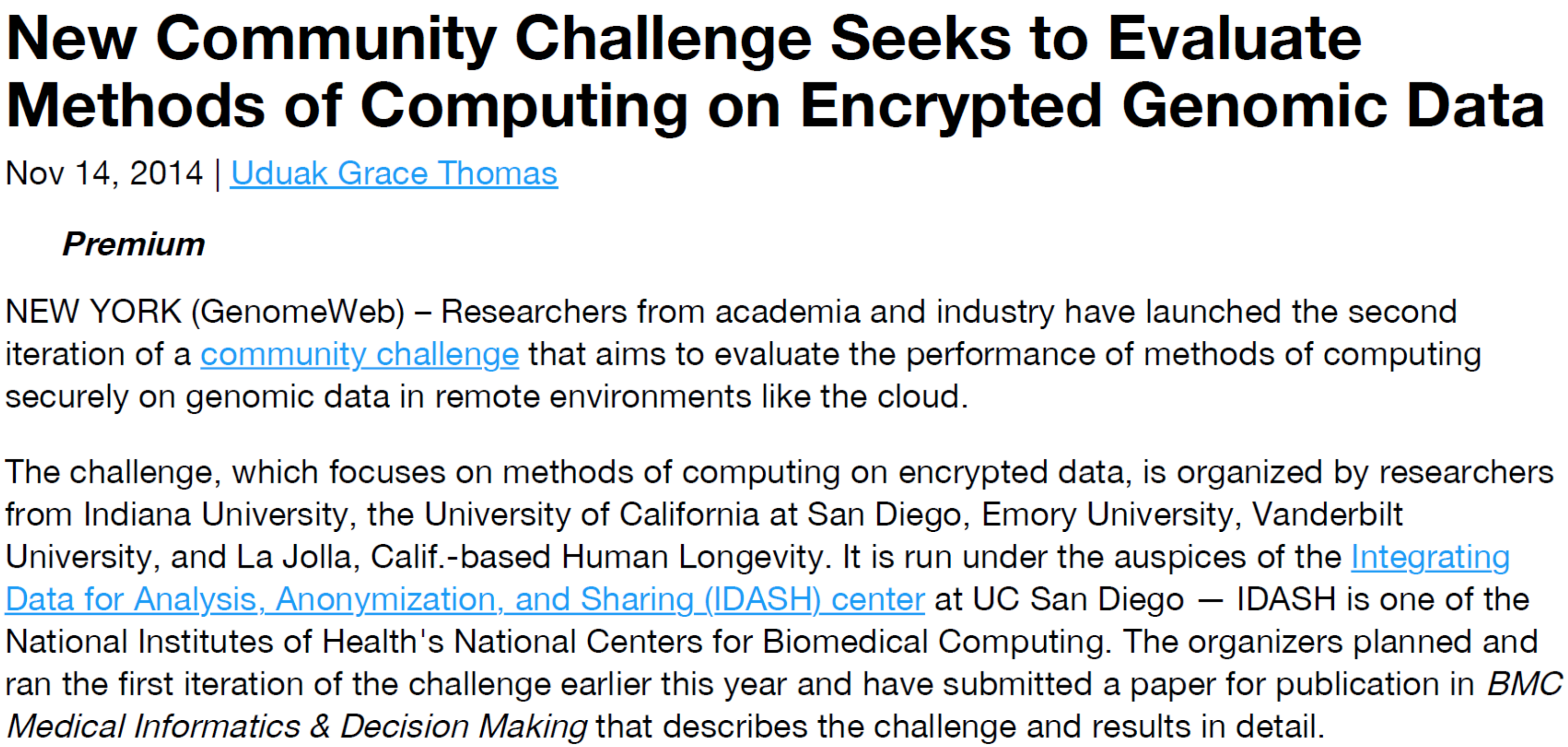
By M Naveed, E. Ayday, E. Clayton, J. Fellay, C. Gunter, JP Hubaux, B. Malin and X. Wang

Available at http://arxiv.org/pdf/1405.1891v1.pdf

Grand Challenges

How to share genomic data in a way that preserves the privacy of the data donors, without undermining the utility of the data or impeding its convenient dissemination?

How to perform a LARGE-SCALE, PRVIACY-PRESERVING analysis on genomic data, in an untrusted cloud environment or across multiple users?



New Community Challenge Seeks to Evaluate Methods of Computing on Encrypted Genomic Data

Nov 14, 2014 | Uduak Grace Thomas

Premium

NEW YORK (GenomeWeb) – Researchers from academia and industry have launched the second iteration of a community challenge that aims to evaluate the performance of methods of computing securely on genomic data in remote environments like the cloud.

The challenge, which focuses on methods of computing on encrypted data, is organized by researchers from Indiana University, the University of California at San Diego, Emory University, Vanderbilt University, and La Jolla, Calif.-based Human Longevity. It is run under the auspices of the Integrating Data for Analysis, Anonymization, and Sharing (IDASH) center at UC San Diego — IDASH is one of the National Institutes of Health's National Centers for Biomedical Computing. The organizers planned and ran the first iteration of the challenge earlier this year and have submitted a paper for publication in *BMC Medical Informatics & Decision Making* that describes the challenge and results in detail.

Real Study, Real Impacts

- Understand the impacts of secure computing techniques on real-world genome analysis:
    - real human genomic data
    - large scale (involving up to 100K sites)

- Balance privacy protection and practical applicability
    - Goal: sufficiently efficient & minimum controlled privacy risks

Challenges and Tasks

- Challenge 1: Homomorphic Encryption (HME) based Genome Analysis
  - Scenario: analyze **encrypted** DNA data on a commercial cloud (e.g., Amazon)
  - Task 1.1: Secure Genome-Wide Association Studies
  - Task 1.2: DNA sequence comparison (Hamming Distance or *Approximate* Edit Distance)

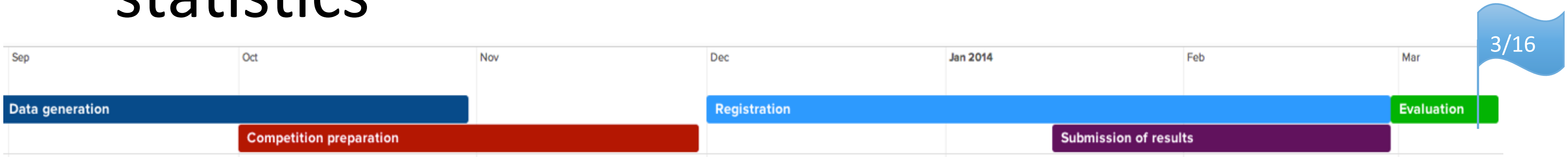
- Challenge 2: Secure Multiparty Computing (SMC) based Genome Analysis
  - Scenario: without exposing their individual data, **two** organizations work together to perform a genomic analysis across their DNA datasets
  - Task 2.1: SMC on GWAS
  - Task 2.2: SMC on sequence comparisons (Hamming and *Approximate* Edit Distances)

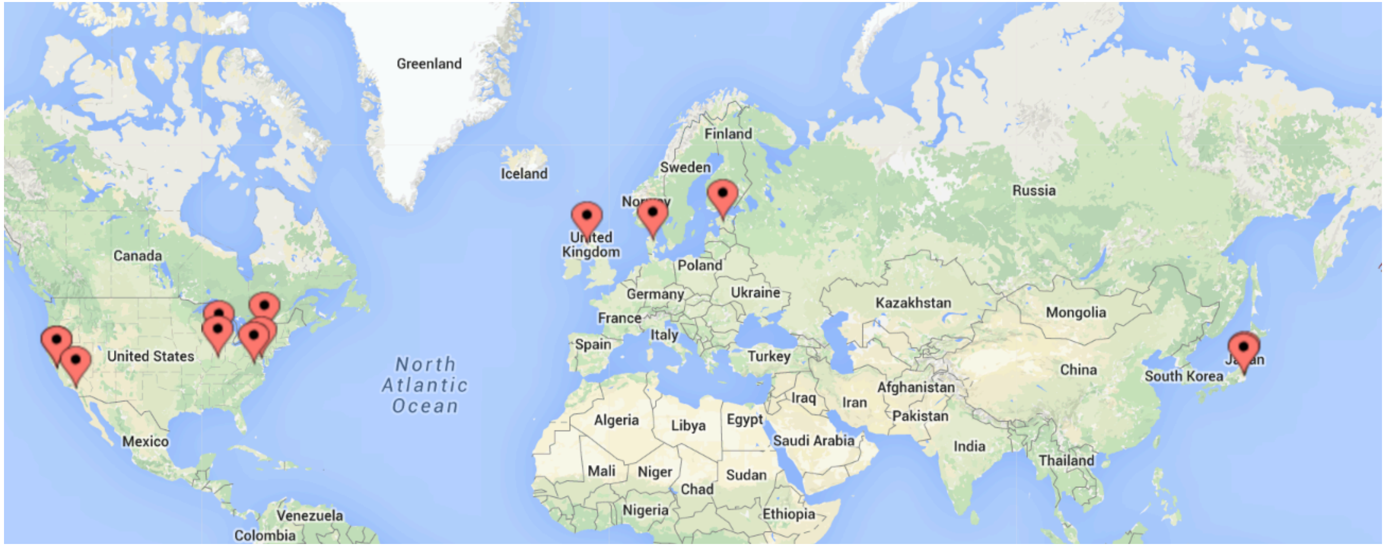
Participant Teams

- 11 Teams, 12 Institutions around the world
  - **North America:** IBM US; Stanford/MIT; Syracuse University; University of Maryland; University of Notre Dame; University of Virginia; Microsoft Research; University of California Irvine;
  - **Europe**: IBM UK; Cybernetica AS (Estonia); The Alexandra Institute (Denmark)
  - **Asia**: University of Tsukuba (Japan)

- Breakdowns across the tasks:
  - Challenge 1: IBM; Stanford/MIT; Microsoft; UCI; University of Tsukuba
  - Challenge 2: Syracuse University; University of Maryland; University of Notre Dame; University of Virginia; UCI; Cybernetica AS; The Alexandra Institute

Workshop preparation and registration statistics





- 5 countries
- 7 states
- 50+ registrations
- Over 1,250 online visits in the last 2 months

Schedule

Morning

- 8:00 am - 8:30 am Breakfast and registration
- 8:30 am - 8:45 am Welcome [Lucila Ohno-Machado]
- 8:45 am - 9:30 am Keynote [Kristin Lauter]
- 9:30 am - 10:10 am Setting the Stage [XiaoFeng Wang, Haixu Tang, Shuang Wang, and Xiaoqian Jiang]
    - Brief presentations of major results for the challenge participants. Discussion will include consideration on how all these approaches are interrelated.
- 10:10 am - 10:20 am Break
- 10:20 am – 12:00 am Session I [Li Xiong]
- 12:00 am - 1:00 pm Networking Lunch

Afternoon

- 01:00pm - 2:00 pm Panel discussion [Bradley Malin]
    - Panel discussion about the emerging privacy challenges in genomic research.
- 2:00 pm – 2:40 pm Session II [Haixu Tang]
- 2:40 pm – 2:45 pm Break
- 2:45 pm – 3:45 pm Session III [Shuang Wang]
- 3:45 pm – 4:00 Award ceremony and Discussion [Amalio Telenti]
    - Present Human Longevity, Inc. sponsored awards. Discuss the plan for the next year challenge.
- 4:00 pm – 4:30 [Xiaoqian Jiang and XiaoFeng Wang] Discussion and next Challenges
- 4:30 pm Adjourn

Setting the Stage

Outline

- Data and Methodology
- Participants and Results
- Discussion

Data and Methodology

Motivations & Tasks

- Addressing two data-intensive computing problems in biomedical research (genome-wide association studies (GWAS) and human genome comparison) under two different scenarios (secure computation outsourcing and secure multiparty computation)

	Gemome-wide association studies (GWAS)	Human genome comparison
Outsourcing	Task 1.1	Task 1.2
Multiparty computation	Task 2.1	Task 2.2

Data Selection

- Data source
    - 200 Cases from Personal Genome Project (PGP: http://www.personalgenomes.org/), missing values filled by using fastPHASE
    - 200 Controls were simulated based on the haplotypes of 174 individuals from CEU population of International HapMap Project (http://hapmap.ncbi.nlm.nih.gov/)
    - 2 individual genomes (hu604D39 with 4,542,542 variations and hu661AD0 with 4,368,847 variations comparing to the reference human genome) were randomly selected from PGP

Genome-wide association studies

Given the genotypes of two groups (representing 200 cases and 200 controls) of individuals over 311/610 SNP sites, participating teams are challenged to come up with secure computing algorithms to compute the minor allele frequencies (MAFs) in each group, and a χ^2 test statistic between the two groups on each site.

Task 1.1: each team is given the genotypes of all cases and controls

Task 2.1: the case and control dataset was horizontally partitioned into two sub-datasets (100 cases and controls in each sub-dataset) distributed to two institutions, where each institution will host a single sub-dataset, and cannot exchange the sub-datasets.

Whole genome comparison

Given the genome sequences (in variant call format, or VCF) from two PGP individuals, participating teams are challenged to come up with secure computing algorithms to compute the hamming distance and edit distance between the genomic sequences.

Task 1.2: each team is given the two genome sequences (in VCF).

Task 2.2: the two genome sequences (in VCF) are distributed to two institutions, where each institution will host a single genome and cannot exchange genomes.

Whole genome comparison

- A subset of variation sites were randomly selected from the >4M sites to form the input data of different size (5K, 10K and 100K were used for final evaluation).
- Hamming distance is computed on the variation sites composed of substitutions in both genomes.
- Edit distance is computed on all given variation sites using approximate algorithm.

Whole genome comparison: edit distance

- Edit distance computation (i.e., following the *N^2* dynamic programming algorithm) is known to be expensive by using secure computing protocols.
  - It takes the SMC protocol (implemented in fastGC) **4.7 hours** to compute the edit distance between two human genomic segments of ~5K nucleotides even on local servers (i.e., no communication overhead).
- We devised an approximation algorithm to compute the edit distance between two human genomic sequences based on their variations from the reference genome sequences (i.e., encoded in the VCF files).
  - It performs well in practice: when applied to the comparison of 20 pairs of human genomic segments of ~5K nucleotides, in 18/20 cases, it reported the exact true edit distance, in 1/20 cases, it reported an approximate distance 1 higher than the true one (28 vs. 27), in 1/20 cases, the approximate distance significantly deviated from the true one (48 vs. 51).
  - This algorithm was recommended to all participation teams of task 1.2 and 2.2 for computing edit distance between two human genome sequences.
  - On the other hand, a different approximation proposed by the IBM team during this competition performed much worse in practice. Out of 20 cases as shown above, in only 5 cases the algorithm reported the exact true edit distance; in 8 cases, the reported edit distance is significantly deviated from the actual one (the largest deviation of 24 vs. 48).

Challenge 1: HME based analysis

- Each participating team is required to develop a homomorphic encryption-based protocol to encrypt these input datasets.
- The encrypted datasets can be used to compute the expected results, i.e., the minor allele frequencies (MAF) and chi-squared statistics for task 1.1, and the Hamming distance and edit distance for task 1.2, on an untrusted remote server.
- The protocol should return the encrypted results (e.g., MAF, χ^2 statistics), which only the data owner with the private key can decrypt.

Challenge 2: SMC based Analysis

- Task 2.1: each participating team is required to develop a distributed cryptographic protocol to securely aggregate the minor allele frequencies (MAF) in two datasets and securely calculate χ^2 statistics for each of the given SNPs.
- Task 2.2: each participating team is required to develop a distributed cryptographic protocol to securely compute the Hamming distance and edit distance between two given human genomes across two institutions.

Submission and Evaluation

- For each task, participating teams are given a testing dataset. Each team should submit a suite of programs to implement their algorithms (either binary executable files or source codes) that should be pre-compiled on given pre-set virtual machines (VMs), where the performance is evaluated by organizers on different datasets.
  - For both tasks of challenge 1, each submitted program was executed within the pre-set virtual machine on a single computer, where the runtime and memory usage were recorded.
  - For both tasks of challenge 2, each submitted program was executed within two virtual machines on two servers located at Indiana University and UCSD, respectively, where the runtime and memory usage on each server and the data size communicated between two servers were recorded. Two submitted programs require a third server in the computation, on which we require minimum computation should be involved.

Participants and Results

Challenge 1: HME based DNA Analysis

Task 1.1 GWAS on encrypted DNA data
Task 1.2 DNA sequence comparison (Hamming, Approximate Edit distances)

5 teams:
IBM; Stanford/MIT; Microsoft; UC Irvine (UCI); University of Tsukuba

Results for Task 1.1: Minor Allele Frequency (training dataset with 311 SNPs)

Teams	Execution Time in seconds					Mem (MB)	Method
	Initialization (e.g., key gen)	Encryption	Evaluation	Decryption	Total		
Microsoft Research	6.51073	10.6353	0.002898	0.292005	17.441	118.08	Helib for BGV scheme (with parameter: p=2, r=9, d=1, c=2, k=80, w=64, L=3, m=5461)
UCI*	0.2006	0.3433	0.008816	0.03589	0.5886	3.320	Hom Paillier Cryto(with parameter:N=1024)
Stanford MIT	0.533		0.041	0.495	1.069	8	HMAC-SHA-256, m=2e32
U of Tsukuba	4.277	14.421	29.164	7.346	55.208	31.808	Helib for BGV scheme (with parameter:p=200003, r=1, d=1, c=3, k=128, w=64, L=3, m=8192)

Results for Task 1.1: Minor Allele Frequency (training dataset with 311 SNPs)

Teams	Execution Time in seconds					Mem (MB)	Method
	Initialization (e.g., key gen)	Encryption	Evaluation	Decryption	Total		
Microsoft Research	6.51073	10.6353	0.002898	0.292005	17.441	118.08	Helib for BGV scheme (with parameter: p=2, r=9, d=1, c=2, k=80, w=64, L=3, m=5461)
UCI*	0.2006	0.3433	0.008816	0.03589	0.5886	3.320	Hom Paillier Cryto(with parameter:N=1024)
Stanford MIT	0.533		0.041	0.495	1.069	8	HMAC-SHA-256, m=2e32
U of Tsukuba	4.277	14.421	29.164	7.346	55.208	31.808	Helib for BGV scheme (with parameter:p=200003, r=1, d=1, c=3, k=128, w=64, L=3, m=8192)

*The algorithm encrypts locals count instead of input data for secure data outsourcing, and was not considered in the competition.

Results for Task 1.1: Minor Allele Frequency (training dataset with 311 SNPs)

Teams	Execution Time in seconds					Mem (MB)	Method
	Initialization (e.g., key gen)	Encryption	Evaluation	Decryption	Total		
Microsoft Research	6.51073	10.6353	0.002898	0.292005	17.441	118.08	Helib for BGV scheme (with parameter: p=2, r=9, d=1, c=2, k=80, w=64, L=3, m=5461)
UCI*	0.2006	0.3433	0.008816	0.03589	0.5886	3.320	Hom Paillier Cryto(with parameter:N=1024)
Stanford MIT	0.533		0.041	0.495	1.069	8	HMAC-SHA-256, m=2e32
U of Tsukuba	4.277	14.421	29.164	7.346	55.208	31.808	Helib for BGV scheme (with parameter:p=200003, r=1, d=1, c=3, k=128, w=64, L=3, m=8192)

*The algorithm encrypts locals count instead of input data for secure data outsourcing, and was not considered in the competition.

Results for Task 1.1: Minor Allele Frequency (testing dataset with 610 SNPs)

Teams	Execution Time in seconds					Mem (MB)	Method
	Initialization (e.g., key gen)	Encryption	Evaluation	Decryption	Total		
Microsoft Research	11.2287	14.3732	0.004673	0.7	26.306	234.72	Helib (with parameter: p=2, r=9, d=1, c=2, k=80, w=64, L=3, m=8191)
UCI*	0.2007	0.6139	0.0114	0.059823	0.8858	3.320	Hom Paillier Cryto(with parameter:N=1024)
Stanford MIT	0.911		0.044	0.892	1.847	13	HMAC-SHA-256, m=2e32
U of Tsukuba	4.186	29.270	64.014	14.853	112.32	32.668	Helib (with parameter:p=200003, r=1, d=1, c=3, k=128, w=64, L=3, m=8192)

Results for Task 1.1: Chi-square statistics (training dataset with 311 SNPs)

Teams	Execution Time in seconds					Mem (MB)	Method
	Initialization (e.g., key gen)	Encryption	Evaluation	Decryption	Total		
Microsoft Research	5.919	10.6529	0.002277	0.301718	16.8759	118.1	Helib (with parameter: p=2, r=10, d=1, c=2, k=80, w=64, L=3, m=5461)
UCI	0.2006	0.3433	0.08816	0.026571	0.6586	3.320	Hom Paillier Cryto(with parameter:N=1024)
Stanford MIT	0.533		0.041	0.495	1.069	8	HMAC-SHA-256, m=2e32
U of Tsukuba	4.277	14.421	29.164	7.346	55.208	31.808	Helib (with parameter:p=200003, r=1, d=1, c=3, k=128, w=64, L=3, m=8192)

Results for Task 1.1: Chi-square statistics (testing dataset with 610 SNPs)

Teams	Execution Time in seconds					Mem (MB)	Method
	Initialization (e.g., key gen)	Encryption	Evaluation	Decryption	Total		
Microsoft Research	11.2756	15.1456	0.004161	0.687744	27.1131	234.73	Helib (with parameter: p=2, r=10, d=1, c=2, k=80, w=64, L=3, m=8191)
UCI	0.2007	0.6139	0.0114	0.04481	0.87081	3.320	Hom Paillier Cryto(with parameter:N=1024)
Stanford MIT	0.911		0.044	0.892	1.847	13	HMAC-SHA-256, m=2e32
U of Tsukuba	4.186	29.270	64.014	14.853	112.323	32.668	Helib (with parameter:p=200003, r=1, d=1, c=3, k=128, w=64, L=3, m=8192)

Result Summary for Task 1.1

	MAF		Chi-square		
	311 SNPs	610 SNPs	311 SNPs	610 SNPs	Time(SEc.)
Microsoft Research	17.4409331	26.306573	16.875895	27.1131054	
UCI*	0.5886	0.8858	0.6586	0.87081	
Stanford/MIT	1.069	1.847	1.069	1.847	
U of Tsukuba	55.208	112.323	55.208	112.323	
	311 SNPs	610 SNPs	311 SNPs	610 SNPs	MemoRy(MB)
Microsoft Research	130.484	247.296	118.080	234.728	
UCI*	3.320	3.320	3.320	3.320	
Stanford/MIT	8.0	13.0	8.0	13.0	
U of Tsukuba	31.808	32.668	31.808	32.668	

*The algorithm encrypts local counts instead of input data for secure data outsourcing, and was not considered in the competition.

Results for Task 1.2 (Hamming)

ACCRACY

	Training		Testing		
	5k	100k	5k	10k	100k
Plaintext data	4740	131535	3099	3306	134252
IBM	4740	131545	3099	3306	134260
Microsoft	4740	N/A	3099	3306	N/A
Stanford/MIT	4720	130035	3082	3275	132703

TIME

	5k	100k	5k	10k	100k
Plaintext data	0.095s	1.274s	0.076s	0.118s	1.145s
IBM	79.0s	475.2s	79.4s	86.8s	472.2s
Microsoft	44.019s	N/A	44.664s	80.031s	N/A
Stanford/MIT	20m25s	1h54m11s	20m37s	36m27s	2h2m26s

MEMORY

	5k	100k	5k	10k	100k
Plaintext data	2.43M	13.52M	1.64M	2.43M	13.52M
IBM	1.416G	2.165G	1.416G	1.419G	2.168G
Microsoft	513.5M	N/A	513.7M	720.5M	N/A
Stanford/MIT	2.765G	7.489G	2.765G	4.025g	7.502G

Teams	Method
IBM	Helib 5K:p=653,r=1,d=2,b=25,c=4,k=86.87, L=19,m=17767 10K:p=653,r=1,d=2,c=4,k=86.8699, b=25, L=19,m=17767 100K:p=653,r=1,d=2,c=4,k=86.8699,b=25, L=19,m=17767
Microsoft	Helib: 5K: p=2, r=1, d=1, c=2, k=80, w=64, L=7, m=8191 10K: p=2, r=1, d=1, c=2, k=80, w=64, L=7, m=8191
Stanford/ MIT	Helib for BGV encryption scheme: p=19259, m=19258, phi(m)=9629, k=80 Hashing: HMAC-SHA-256 5K: k=1000000 b=1 m=3 10K: k=1700000 b=1 m=3 100K: k=5000000 b=1 m=3

Results for Task 1.2 (Hamming)

	Training		Testing			
	5k	100k	5k	10k	100k	
Plaintext data	4740	131535	3099	3306	134252	ACCRACY
IBM	4740	131545	3099	3306	134260	
Microsoft	4740	N/A	3099	3306	N/A	
Stanford/MIT	4720	130035	3082	3275	132703	

	5k	100k	5k	10k	100k	
Plaintext data	0.095s	1.274s	0.076s	0.118s	1.145s	TIME
IBM	79.0s	475.2s	79.4s	86.8s	472.2s	
Microsoft	44.019s	N/A	44.664s	80.031s	N/A	
Stanford/MIT	20m25s	1h54m11s	20m37s	36m27s	2h2m26s	

	5k	100k	5k	10k	100k	
Plaintext data	2.43M	13.52M	1.64M	2.43M	13.52M	MEMORY
IBM	1.416G	2.165G	1.416G	1.419G	2.168G	
Microsoft	513.5M	N/A	513.7M	720.5M	N/A	
Stanford/MIT	2.765G	7.489G	2.765G	4.025g	7.502G	

Teams	Method
IBM	Helib 5K:p=653,r=1,d=2,b=25,c=4,k=86.87, L=19,m=17767 10K:p=653,r=1,d=2,c=4,k=86.8699, b=25, L=19,m=17767 100K:p=653,r=1,d=2,c=4,k=86.8699,b=25, L=19,m=17767
Microsoft	Helib: 5K: p=2, r=1, d=1, c=2, k=80, w=64, L=7, m=8191 10K: p=2, r=1, d=1, c=2, k=80, w=64, L=7, m=8191
Stanford/ MIT	Helib for BGV encryption scheme: p=19259, m=19258, phi(m)=9629, k=80 Hashing: HMAC-SHA-256 5K: k=1000000 b=1 m=3 10K: k=1700000 b=1 m=3 100K: k=5000000 b=1 m=3

Results for Task 1.2 (Hamming)

	Training		Testing			
	5k	100k	5k	10k	100k	ACCRACY
Plaintext data	4740	131535	3099	3306	134252	
IBM	4740	131545	3099	3306	134260	
Microsoft	4740	N/A	3099	3306	N/A	
Stanford/MIT	4720	130035	3082	3275	132703	
	5k	100k	5k	10k	100k	TIME
Plaintext data	0.095s	1.274s	0.076s	0.118s	1.145s	
IBM	79.0s	475.2s	79.4s	86.8s	472.2s	
Microsoft	44.019s	N/A	44.664s	80.031s	N/A	
Stanford/MIT	20m25s	1h54m11s	20m37s	36m27s	2h2m26s	
	5k	100k	5k	10k	100k	MEMORY
Plaintext data	2.43M	13.52M	1.64M	2.43M	13.52M	
IBM	1.416G	2.165G	1.416G	1.419G	2.168G	
Microsoft	513.5M	N/A	513.7M	720.5M	N/A	
Stanford/MIT	2.765G	7.489G	2.765G	4.025g	7.502G	

Teams	Method
IBM	Helib 5K:p=653,r=1,d=2,b=25,c=4,k=86.87, L=19,m=17767 10K:p=653,r=1,d=2,c=4,k=86.8699, b=25, L=19,m=17767 100K:p=653,r=1,d=2,c=4,k=86.8699,b=25, L=19,m=17767
Microsoft	Helib: 5K: p=2, r=1, d=1, c=2, k=80, w=64, L=7, m=8191 10K: p=2, r=1, d=1, c=2, k=80, w=64, L=7, m=8191
Stanford/ MIT	Helib for BGV encryption scheme: p=19259, m=19258, phi(m)=9629, k=80 Hashing: HMAC-SHA-256 5K: k=1000000 b=1 m=3 10K: k=1700000 b=1 m=3 100K: k=5000000 b=1 m=3

Results for Task 1.2 (Approximate Edit distances)

	Training		Testing			
	5k	100k	5k	10k	100k	ACCURACY
Plaintext data	7446	198705	9089	16667	191986	
IBM*	5777	153266	5328	8318	153266	
Microsoft	7446	N/A	9089	16665	N/A	
	5k	100k	5k	10k	100k	TIME
Plaintext data	0.103s	1.489s	0.106s	0.144s	1.528s	
IBM*	96.9s	552.6s	91.7s	106.3s	555.2s	
Microsoft	92.26s	N/A	91.09s	181.92s	N/A	
	5k	100k	5k	10k	100k	MEMORY
Plaintext data	2.45M	25.78M	2.45M	2.53M	25.78M	
IBM*	1.416G	2.294G	1.418G	1.451G	2.295G	
Microsoft	701.1M	N/A	700.8M	1.295G	N/A	

Teams	Method
IBM	Helib 5K:p=653,r=1,d=2,b=25,c=4,k=86.87, L=19,m=17767 10K:p=653,r=1,d=2,c=4,k=86.8699, b=25, L=19,m=17767 100K:p=653,r=1,d=2,c=4,k=86.8699, b=25, L=19,m=17767
Microsoft	Helib 5K : p=2, r=1, d=1, c=2, k=80, w=64, L=9, m=8191 10K: p=2, r=1, d=1, c=2, k=80, w=64, L=11, m=8191

*An approximate algorithm (with about 22% error), which was not considered in the competition.

Winners

Task 1.1: Stanford/MIT

Task 1.2: Hamming distance: IBM

Task 1.2: Approximate Edit distance: Microsoft

Challenge 2: Secure Collaboration on DNA Analysis

Task 1.1 Two-party Privacy-Preserving GWAS
Task 1.2 Two-Party DNA comparison (Hamming,Edit distances)

7 teams:
Syracuse University (SU); University of Maryland (UMD); University of Notre Dame (UND); University of Virginia (UV); UC Irvine (UCI); Cybernetica AS (CAS); The Alexandra Institute (AI)

Results for Task 2.1: χ^2-statistics (small dataset with 311 SNPs)

	Time(s)	Memory (KB)			Communication (MB)		
		VM1	VM2	VM3	VM1	VM2	VM3
Baseline	92	1.2	1.4		0.7	35.0	
UV	32	3.3	5.3		1.9	163.0	
UND	15	25.1	25.1	25.0	4.0	3.8	3.8
SU	14*	173K	162K		4942.4	45.6	
UMD	13	63.5	58.1		0.8	46.2	
CAS	60	0.1	0.1	0.1	0.007	0.007	0.007

* Updated results on April 2

Results for Task 2.1: χ^2-statistics (large dataset with 610 SNPs)

	Time(s)	Memory (KB)			Communication (MB)		
		VM1	VM2	VM3	VM1	VM2	VM3
Baseline	187	1.2	1.4		1.4	70.0	
UV	59	6.9	9.7		3.6	309.3	
UND	23	36.2	49.8	36.0	7.9	7.4	7.2
SU	54*	187	175		9645.7	93.0	
UMD	20	71.3	64.6		1.6	90.7	
CAS	57	0.1	0.1	0.1	0.007	0.007	0.007

* Updated results on April 2

Results for Task 2.2: Hamming Distance (over ~100K variation sites)

	Time(s)	Memory(MB)			Communication(MB)		
		VM1	VM2	VM3	VM1	VM2	VM3
UV	553	0.3	0.3		156.5	9672.9	
UND	5077	3044	3048	3048	4118.5	3361.7	3167.3
UMD	604	1260	1252		63.4	2973.3	
UMD (BF)**	83	0.1	0.1		19.8	150.8	
UCI	788	0.4	0.4		28.8	24.4	
CAS*	128	0.4	0.4	0.4	0.1	0.1	0.1

*The algorithm involves intensive computation on the third server, and thus was not considered in the competition.

**An approximate algorithm (with about 0.8% error) based on Bloom filter, which was not considered in the competition.

Results for Task 2.2: Edit Distance (over ~100K variation sites)

	Time(s)	Memory(KB)			Communication(MB)		
		VM1	VM2	VM3	VM1	VM2	VM3
Baseline	254	290	292		92.0	5595.0	
UMD	>20h						
UMD (BF)**	233	145	125		50.2	424.5	
UCI	998	434	398		39.1	32.7	
AI	>20h						

**An approximate algorithm (with about 0.8% error) based on Bloom filter, which was not considered in the competition.

Winners

Task 2.1: University of Maryland

Task 2.2, Hamming distance: University of Virginia

Task 2.2, Edit distance: University of California, Irvine

Secure DNA Analysis:
Where We Stand?

Moving Closer to Practical Use

- Analyzing Encrypted DNA
  - Hamming and Edit distance approximation over 100K can be done within 10 minutes
- Secure collaboration across the Internet
  - χ^2 based GWAS over hundreds of SNPs can be done, securely, in a few minutes
  - Hamming distance can be calculated in 10 minutes and Edit distance in 20 minutes over 100K across the Internet (Indiana to San Diego)
- We are really close to protecting Some DNA analyses at a practical scale

But Still not There, Yet

- A full-fledged GWAS still cannot be efficiently done on encrypted DNA
    - Due to the challenge of performing divisions efficiently

- HME needs multi-gigabytes of memory and SMC needs to transmit multi-gigabytes of data across the Internet, for analyzing a 100K sequence

- Operations that can be conducted in seconds can take a dozen minutes or hours to compute

- Accurate edit distance is still off the table

How to Bridge the Gap

- Make crypto primitives faster, more lightweight
- Specialize protection for DNA analysis
  - E.g., somewhat HE works better than FHE
- Approximate complicated computation
  - convert a hard-to-protect analysis to those that can be done
- Partition the computing tasks
  - customize computation based on the feature of the problem

Acknowledgements

- Funding:
  - iDASH (U54HL108460)
  - NCBC-collaborating R01 (HG007078-01)
  - NLM (R00LM011392, R21LM012060)
  - NHGRI (K99HG008175)
- Student evaluators
  - Special thanks to Yongan Zhao, Yuchen Zhang, Wenrui Dai, Yong Li
- Administrative support
  - Special thanks to Rita Germann-Kurtz, Morgan Von Ebke

Summary and Discussion

Take home message

- We are making progress on large-scale, secure computing of real-world genomic analysis tasks, but the gap is still there

- Narrowing the gap needs a **Joint** effort from the folks in bioinformatics, biomedic, cryptography, system security and bioethics

- The key here is to connect what cryptography can do and what genomic research and applications need to do
  - E.g., a new infrastructure gathers the most effective/efficient crypto primitives to build the services that biomedical/bioinformatics researchers and practitioners use

Follow-up

- BMC special issue on Human Genome Privacy
- Every participant is encouraged to submit a paper
- There is a publication fee involved
- Timeline

Next Competition

- Secure computations on other biomedical/bioinformatics tasks?

- What secret-sharing based approaches can achieve? What are the legal implications of their assumptions?

- How about those “good-enough” security techniques? Are they practical enough and indeed good enough?

- Protect data and prevent inference?

2nd Workshop on Genome Privacy (GenoPri): May 21, San Jose

- A forum for discussing state-of-the-art genome privacy technologies
  - Informal publication, discussion style

- Example topics:
  - Privacy preserving genome-data analysis and dissemination, access control on genomic data, crypto techniques designed for genome protection, genome privacy with family members, storage protection of genomic data, etc.