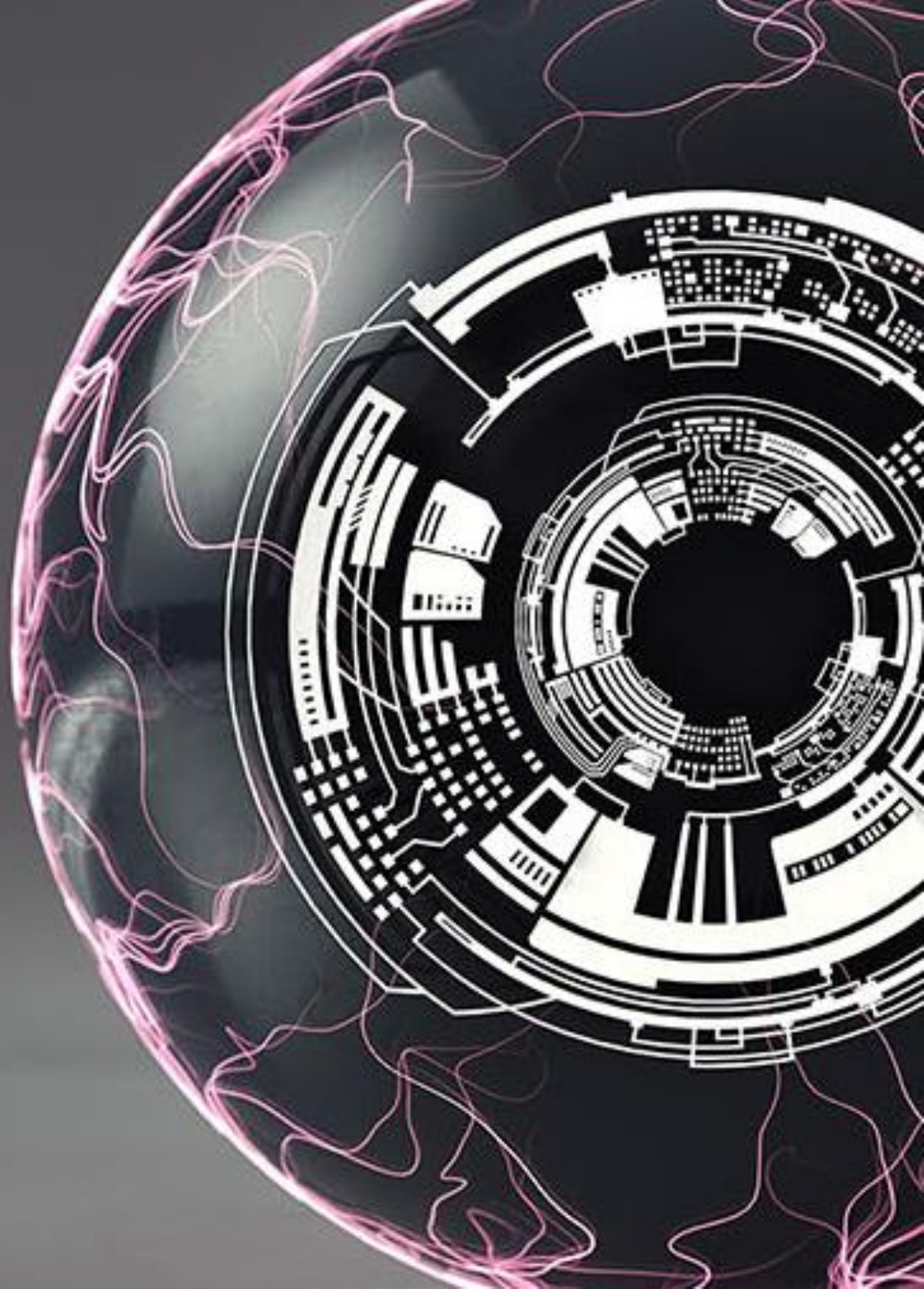
Microsoft

iDASH Privacy and Security workshops

Xiaoqian Jiang, UCSD



Human Genome Privacy

Human DNA is important to genomic research, biomedical studies, and is becoming part of electronic health records (EHRs)

Examples: Genome-wide association studies (GWAS), rare disease studies, targeted therapy, precision medicine

However, genomic data are also highly sensitive

Personally identifiable markers: skin, hair color, predisposition to disease…

Examples of breach: Disease markers, surname identification, face reconstruction

Grand Challenge

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

Utility and Privacy Balance

Secure primitives increase the computational cost, noise adding brings in artifacts to human genome data, there is a critical tradeoff

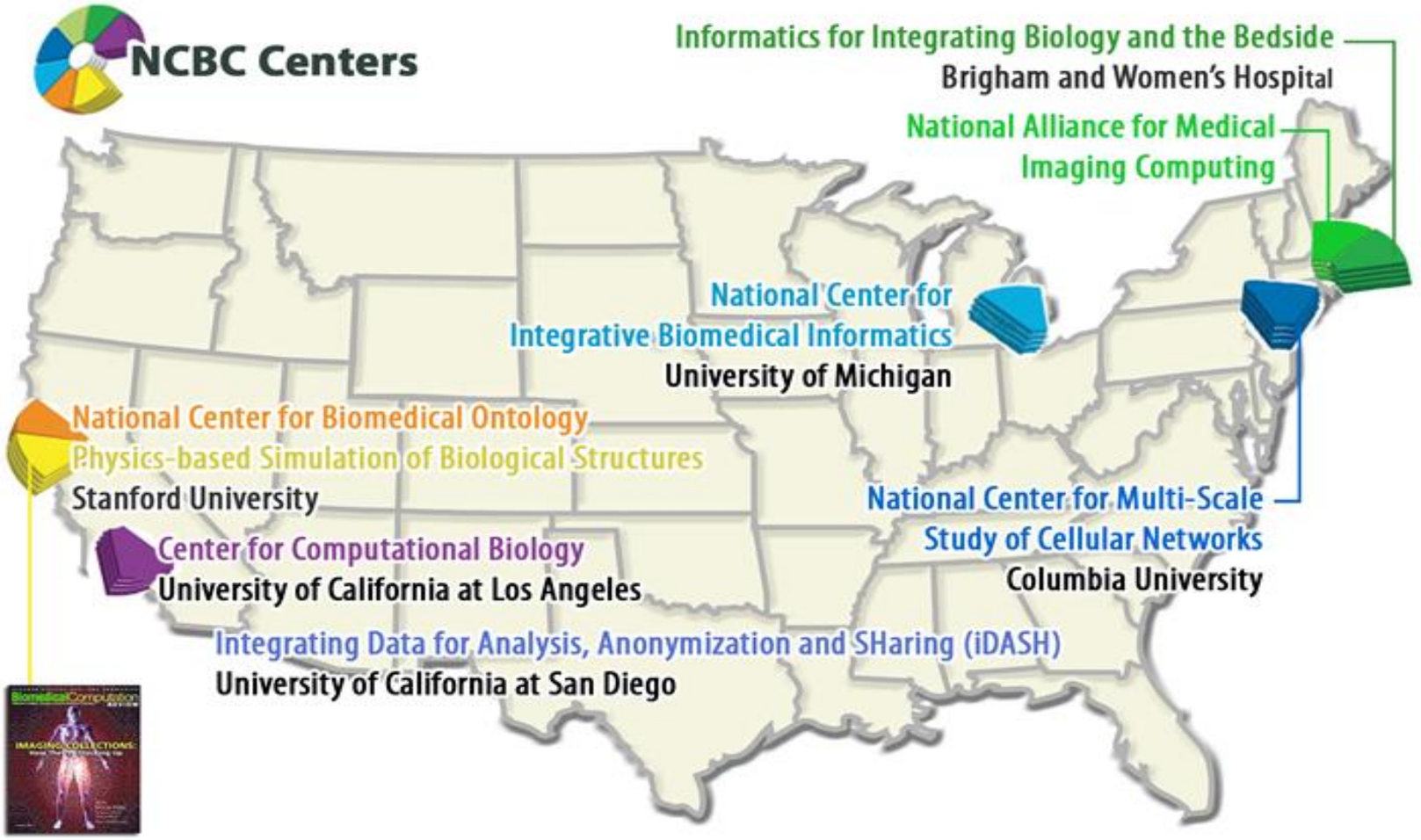
Questions: Can state-of-the-art techniques be used to support biomedical research in practice?

NIH Roadmap

National Centers for Biomedical Computing

Home | NCBC Summary | Calendar | All Hands Meetings | Biological Projects | Biositemaps Projects | Working Group Archive

Search for NCBC resources in the new Resource Discovery System (RDS)



Supported by the National Institutes of Health through the NIH Roadmap for Medical Research, Grant U54HL108460, PI: Ohno-Machado

Real Study, Real Impacts

Understand the impacts of data “anonymization” and secure models to real-world biomedical studies

Real human genomic data

High dimension of a practical scale

Balance privacy/security protection and utility

Goal: maximum utility with minimum controlled privacy risks

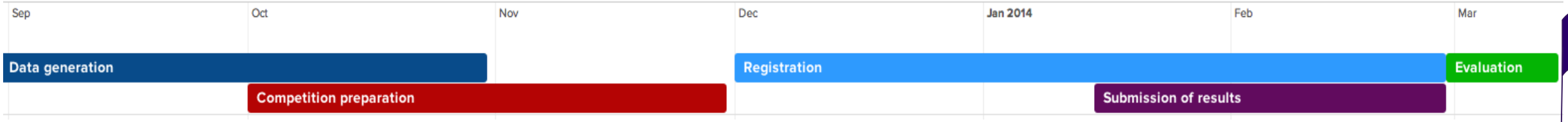
1st iDASH S&P competition (2014)

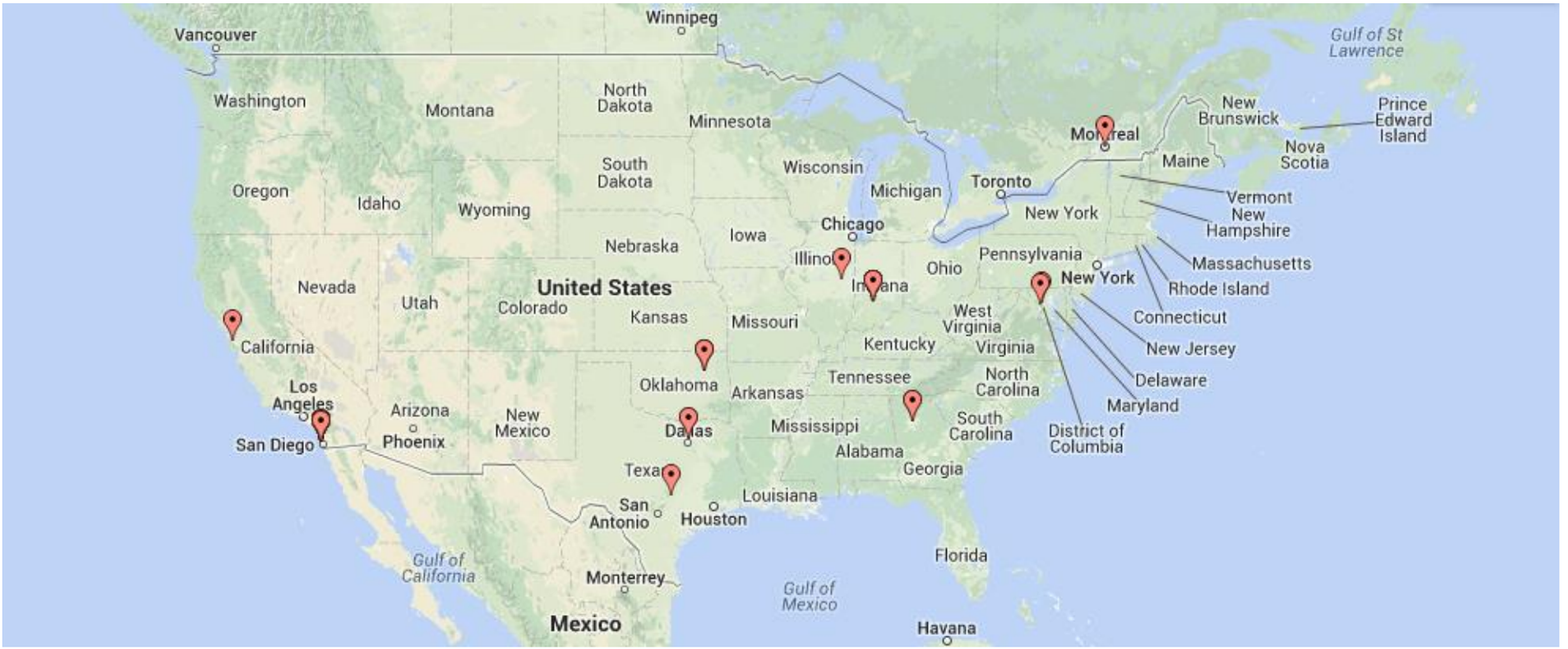
The 1st Competition

Evaluate how effective the best privacy protecting technologies could be in protecting human genomic data and analysis results

The challenge focused on tasks related to sharing aggregate SNP data (allele frequencies) and top-K SNP identification for GWAS studies

Workshop preparation and registration statistics





- 2 countries
- 9 states
- 33 registrations

Challenge of Task 1

Goal: Understand the privacy-utility balance achievable when publicly released SNP data, after proper 'anonymization,' for a realistic GWAS

Utility: the number of significant SNPs identified by the Chi-square association test over the case population (200 individuals from PGP) and a control population (from HapMap)

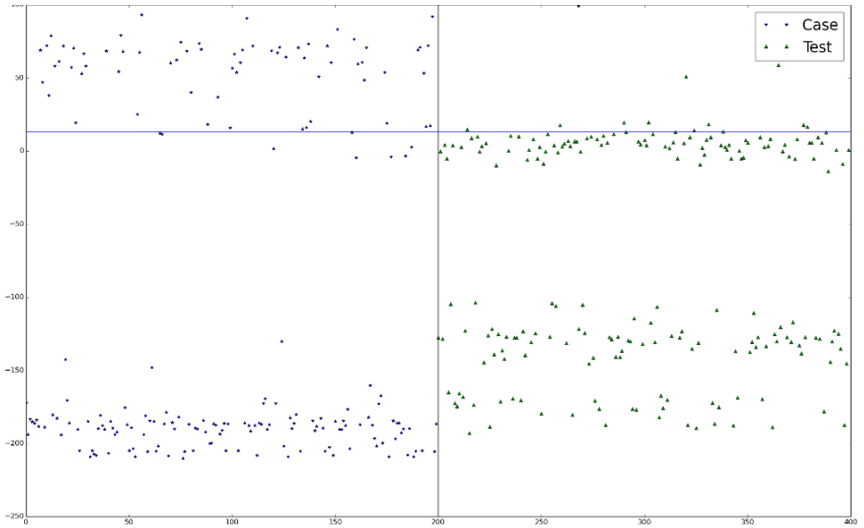
Privacy Protection: the 'anonymized' data's resistance to one of the strongest re-identification statistical attack (i.e., the likelihood ratio test).

Sankararaman, S., Obozinski, G., Jordan, M. I., & Halperin, E. (2009). *Nature Genetics*, *41*(9), 965–7. doi:10.1038/ng.436

Privacy: Evaluation of Privacy Risks using the Likelihood Ratio Test

$$\bar{L} = \sum_{j}^{m} \left(x_j \log \frac{\hat{p}_j}{p_j} + (1 - x_j) \log \frac{1 - \hat{p}_j}{1 - p_j} \right)$$

where m is the number of SNPs, p_j is the allele frequency of SNP j in the population and $\hat{p}_j$ is that in a pool



Implemented as an online tool that allows challenge participants to examine privacy risks in their noise-added data: http://humangenomeprivacy.org

Sankararaman, S., Obozinski, G., Jordan, M. I., & Halperin, E. (2009). *Nature Genetics*, *41*(9), 965–7. doi:10.1038/ng.436

Utility: Case-Control Association Test

Chi-square: $\chi^2 = \sum_{i=1}\sum_{j=1} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$

$O_{i,j}$ is observed frequencies, $E_{i,j}$ is expected frequencies

Observed allele counts for SNP1

SNP1	A	T	Total
Case	a = 3	b = 1	r = 4
Control	c = 1	d = 3	s = 4
Total	a + c	b+d	n = 8

Expected allele counts for SNP1

A	T
(a+c)*r/n	(b+d)*r/n
(a+c)*s/n	(b+d)*s/n

Challenge of Task 2

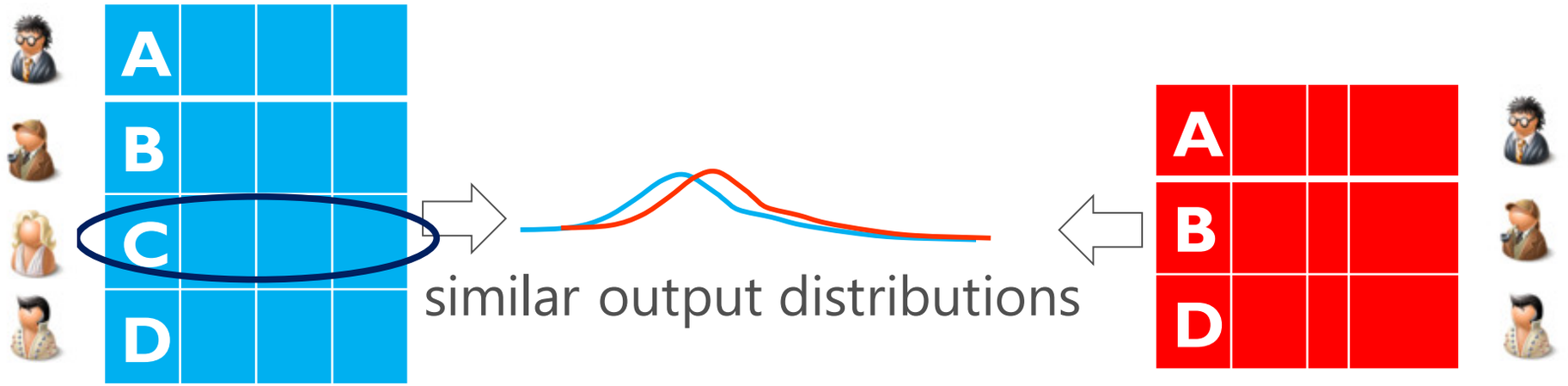
Goal: Given a privacy protection standard, evaluate how much utility, in terms of top-*K* most significant SNPs, can be preserved by the best techniques for 'anonymized' outcome release

Utility: Top-*K* most significant SNPs (using chi-square tests) across the genome (e.g., *K*=1 or 5)

Privacy Protection: Differential privacy with a budget ε=1.0

Differential Privacy

A mechanism is differentially private if every output is produced with similar probability whether any given input is included or not



Risk for C does not increase much if her data are included in the computation

C. Dwork, "Differential privacy," *Int. Colloq. Autom. Lang. Program.*, vol. 4052, no. d, pp. 1–12, 2006.

Data preparation

Personal genome project

CEU population in HapMap

Filtered and genotyped

Task 1: data publishing

Case: 200 PGP individuals
Control: 174 CEU individuals
Data set 1: 311 SNVs
Data set 2: 610 SNVs

Task 2: top-K SNP identification

Case: 200 PGP individuals
Control: 174 CEU individuals
Data set 1: 5000 SNVs
Data set 2: 106,129 SNVs

Task 1: Privacy Preserving Data Sharing

		Baseline		Team 1	Team 2	Team 3	# of sig SNVs
		SNV-based	Haplotype-based	U Oklahoma	UT Dallas	McGill U	
	Power	0.05	0.03	0.61	0.04	0.01	
	Cutoff	TPR. FPR	TPR. FPR	TPR. FPR	TPR. FPR	TPR. FPR	
D1	5.00E-02	0.864. 0.844	0.910. 0.612	1.000. 0.941	1.000. 0.855	1.000. 0.886	22
	1.00E-03	0.632. 0.774	1.000. 0.493	1.000. 0.884	1.000. 0.791	1.000. 0.798	19
	1.00E-05	0.643. 0.700	1.000. 0.475	1.000. 0.879	1.000. 0.737	1.000. 0.737	14
	Power	0.04	0.115	0.005	0.01	0.09	
	Cutoff	TPR. FPR	TPR. FPR	TPR. FPR	TPR. FPR	TPR. FPR	
D2	5.00E-02	0.933. 0.924	0.978. 0.804	1.000. 0.958	0.533. 0	0.956. 0.746	45
	1.00E-03	0.800. 0.862	1.000. 0.708	1.000. 0.909	1.000. 0	1.000. 0.582	15
	1.00E-05	0.625. 0.788	1.000. 0.504	1.000. 0.876	1.000. 0	1.000. 0.425	8

In the first column, D1 refers to 200 participants, 311 SNVs (~29504091-30044866, chr2) and D2 refers to 200 participants, 610 SNVs (~55127312-56292137, chr10). The rows labeled 'Power' indicate the ratio of identifiable individuals using the likelihood ratio test in the case group. The other rows start with a cutoff threshold for the χ^2test (e.g., 5×10^{-2} ,10^{-3}, 10^{-5}), for which two measurements (true positive rate and false positive rate for SNVs using the χ^2 test) were calculated under each method. The last column corresponds to the number of significant SNVs ($p=10^{-5}$) calculated using the original data (i.e., without added noise).

Task 2: Privacy Preserving Feature Selection

	Teams	Top 1	Top 3	Top 5	Top 10	Top 15	Top 20	Top 30
Small (5000 SNVs)	UT Austin	1	2.66	4.44	8.48	7.07	4.68	2.37
	CMU	0.98	2.28	3.53	7.89	4.59	2.32	1.16
Large (100K SNVs)	UT Austin	1	2.65	4.41	5.90	2.26	0.69	0.18
	CMU	0.98	2.26	3.56	3.27	0.42	0.15	0.07

The table shows the average number of (1000 iterations) privacy-preserving SNV identification algorithms developed by the two participating teams. Both algorithms were trained using the small dataset consisting of 5000 SNVs, and then were tested on both small and large datasets, i.e., select top K (i.e., K = 1, 3, ..., 30) most significant SNVs.

Discussion

It remains a challenge to do privacy-preserving sharing of aggregate human genomic data while maintaining utility in genome-wide association studies (GWAS)

Even for a single genomic locus involving a few hundreds of SNPs, the utility of the data was largely damaged after noise was added to ensure privacy protection

It is unlikely that current privacy-preserving techniques will scale well for sharing whole human genomic data

Discussion

Privacy-preserving techniques show promise on publishing outcomes of GWAS-like analyses

High accuracy can be achieved when only a small number of most significant SNPs are disclosed (from the users' perspective)

This is aligned with a data computing model that only releases the results of analyses to users

2nd iDASH S&P competition (2015)

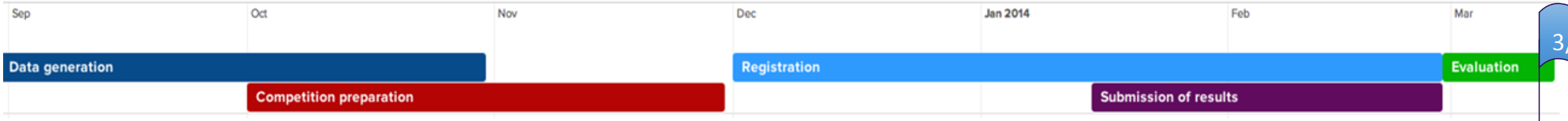
Secure genome analysis competition

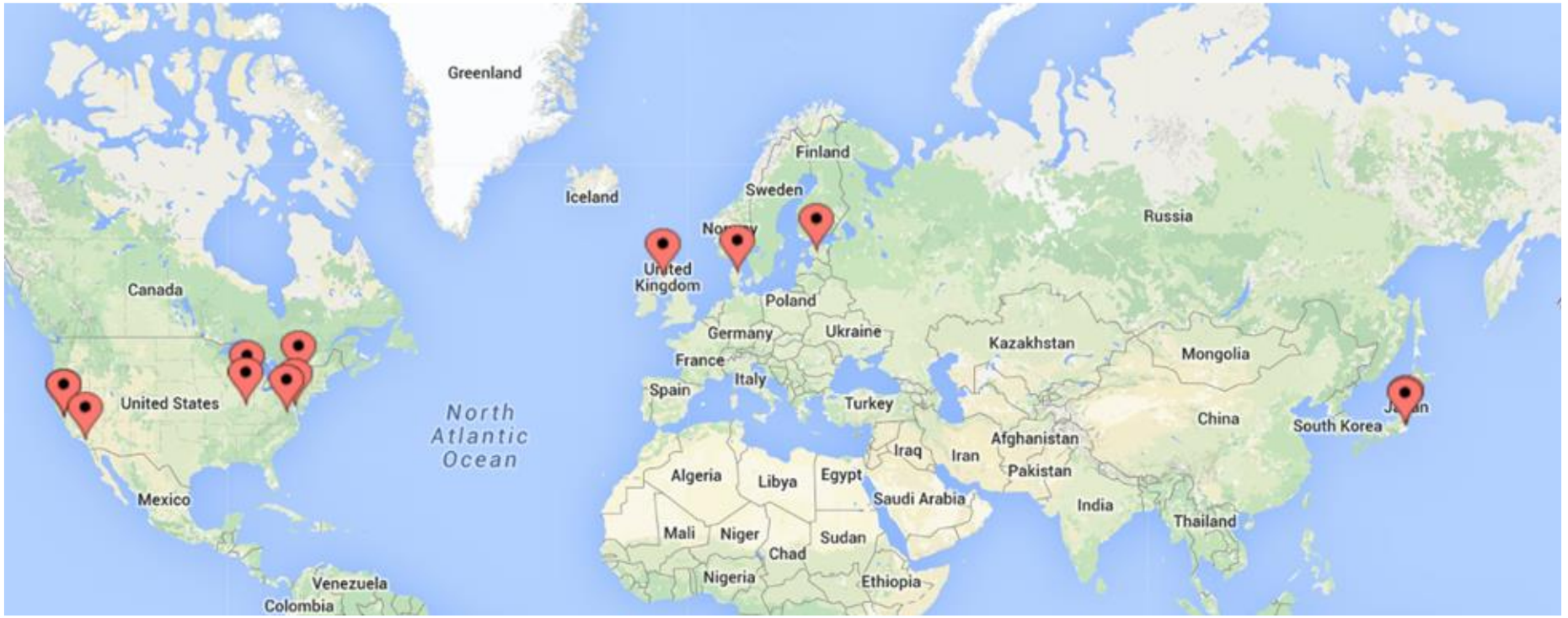
Foster research to address secure outsourcing and multiparty collaboration in biomedical studies

Secure Genome-Wide Association Study (GWAS)

Secure genome comparison based on Hamming and Edit distances

Workshop preparation and registration statistics





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

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Auth

Archive > Volume 519 > Issue 7544 > News > Article

NATURE | NEWS

عربي

Extreme cryptography paves way to personalized medicine

Encrypted analysis of data in the cloud would allow secure access to sensitive information.

Erika Check Hayden

23 March 2015

PDF | Rights & Permissions

David Paul Morris/Bloomberg via Getty

Cloud processing of DNA sequence data promises to speed up discovery of disease-linked gene variants.

The dream for tomorrow's medicine is to understand the links between DNA and disease — and to tailor therapies accordingly. But scientists working to realize such 'personalized' or 'precision' medicine have a problem: how to keep genetic data and medical records secure while still enabling the massive, cloud-based analyses needed to make meaningful associations. Now, tests of an emerging form of data encryption suggest that the dilemma can be solved.

At a workshop on 16 March hosted by the University of California, San Diego (UCSD), cryptographers analysed test genetic data. Working with small data sets, and using a method known as homomorphic encryption, they could find disease-associated gene variants in about ten minutes. Despite the fact that computers were still kept bogged down for hours by more-realistic tasks — such as finding a disease-linked variant in a stretch of DNA a few hundred-thousandths the size of the whole genome — experts in cryptography were encouraged.

genomeweb

Home | Business & Policy | Technology | Research | Clinical | Disease Areas

Home » The Scan » To Keep It Safe and Sound

To Keep It Safe and Sound

Mar 25, 2015

One of the concerns about using genetic data along with medical records information to personalize medicine is how to keep that personal information safe, but still easily accessible for analysis. Cryptographers at a workshop hosted by the University of California, San Diego, tested a homomorphic encryption method that seems promising, reports *Nature* News' Erika Check Hayden.

This method involves mathematically encrypting data on a local computer and then uploading the encoded form to the cloud where it can be analyzed, Check Hayden notes. Encoded results are then sent back to a local computer, which unscrambles the data. Any data intercepted along the way would be encrypted.

She notes that this idea dates back to 1978, but remained largely theoretical until 2009 when IBM Thomas J. Watson Research Center's Craig Gentry showed that computational analyses could be carried out on homomorphically encrypted data.

At the UCSD workshop, cryptographers showed that such an approach could analyze data from 400 people within about 10 minutes and pinpoint a variant associated with disease from among few hundred loci. Analysis of larger datasets and more base pairs wasn't always possible, Check Hayden says, and it could take a lot of computer memory, time, or money.

While the workshop organizers find the approach promising, others say it might not provide enough protection for the data or allow researchers and clinicians to perform all the analyses they want. US National Center for Biotechnology Information's Steven Sherry, for instance, prefers restricting data access to a select few people who have agreed to follow certain regulations on how the data may be used.

11 Teams
12 Institutions

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)

Microsoft

Challenge 1: HME based analysis

Develop a homomorphic encryption-based protocol to analyze encrypted DNA data on an untrusted cloud

Compute the minor allele frequencies (MAF) and chi-square 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, distance), which only the data owner with the private key can decrypt.

Challenge 2: SMC based Analysis

Assess solutions to enable two parties to work together to perform a genomic analysis across their DNA datasets without exposing their individual data

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 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.

Result Summary for Task 1.1

	MAF		Chi-square		Time (Sec.)
	311 SNPs	610 SNPs	311 SNPs	610 SNPs	
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.

Result for Task 1.2 (Hamming distance)

	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	ACCRACY
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.

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.


HUMAN LONGEVITY, INC.
March 16, 2015
iDASH Genome Privacy Protection Competition Winner
$ 300.00
PAY TO THE ORDER OF
Three Hundred
DOLLARS
iDASH

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 types of DNA analyses at a practical scale

But Still not There, Yet

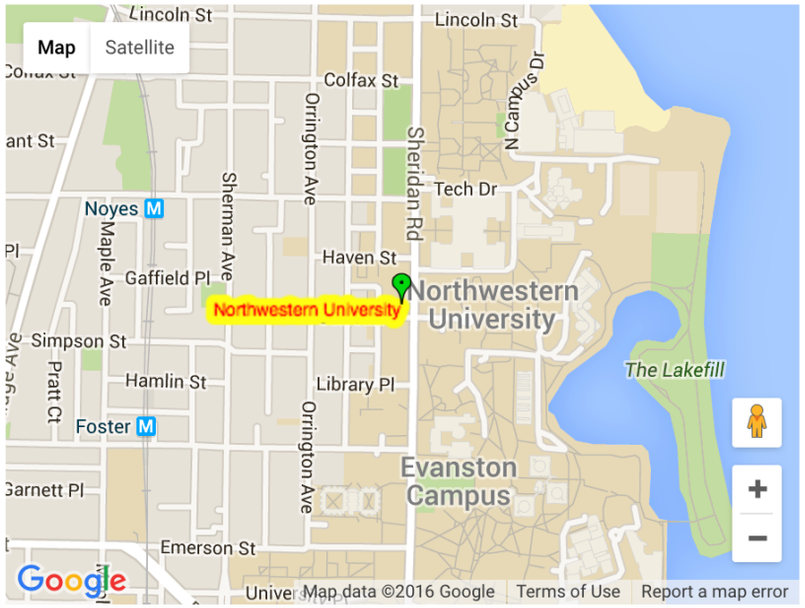
- A full-fledged GWAS still cannot be efficiently done on encrypted DNA
  - Due to the challenge of performing divisions efficiently
- HME needs multi-gigabytes of memory and SMC needs to transmit multi-gigabytes of data across the Internet, for analyzing a 100K sequence
- Operations that can be conducted in seconds can take a dozen minutes or hours to compute
- Accurate edit distance is still off the table

3rd iDASH S&P Competition (2016)

Microsoft Research
Faculty Summit 2016

Workshop preparation and registration statistics

Workshop Location



Registered Teams

- 13 countries
- 50+ teams

Theme of 2016 (humangenomeprivacy.org)

Tackles emerging and practical problems, evaluation will balance performance, security guarantee and, importantly, the generality of the solution

Track 1: Practical Protection of Genomic Data Sharing through Beacon Services (privacy-preserving output release)
Track 2: Privacy-Preserving Search of Similar Cancer Patients across Organizations (secure multiparty computing)
Track 3: Testing for Genetic Diseases on Encrypted Genomes (secure outsourcing)

Competition organizers

- Haixu Tang (Indiana University)
- XiaoFeng Wang (Indiana University)
- Shuang Wang (UCSD)
- Xiaoqian Jiang (UCSD)

Local organizers

- Bradley Malin (Vanderbilt University)
- Abel Kho (Northwestern University)

General chair

- Lucila Ohno-Machado (UCSD)

Acknowledgements

- Xiaofeng Wang (Indiana University)
- Haixu Tang (Indiana University)

- Shuang Wang (UCSD)
- Lucila Ohno-Machado (UCSD)

Supported by the National Institutes of Health through the NIH Roadmap for Medical Research, Grant U54HL108460, NHGRI Grant R01HG007078