IDASH PRIVACY & SECURITY WORKSHOP 2016 -- COMPETITION RESULTS

Competition organizers:

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

THREE TRACKS OF COMPETITION TASKS

- Track 1: Practical Protection of Genomic Data Sharing through Beacon Services (Privacy-preserving data dissemination)
- Track 2: Privacy-Preserving Search of Similar Cancer Patients across Organizations (Secure collaboration)
- Track 3: Testing for Genetic Diseases on Encrypted Genomes (Secure outsourcing)

TRACK 1: PRACTICAL PROTECTION OF GENOMIC DATA SHARING THROUGH BEACON SERVICES

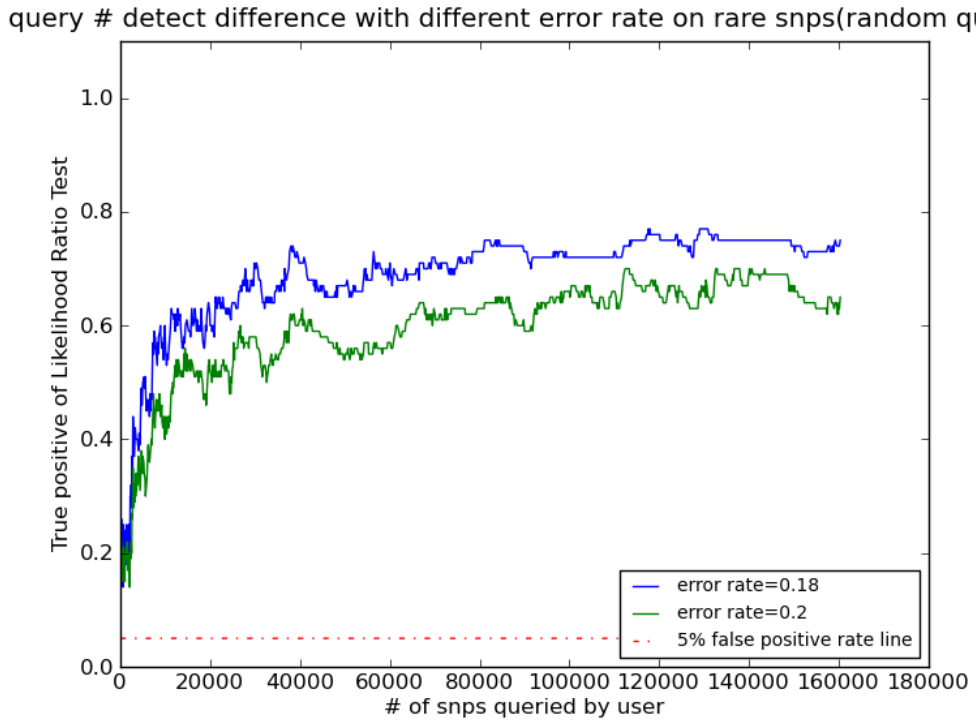
- Background: The **Beacon project** was created by the Global Alliance for Genomics and Health (GA4GH) as a means of "testing the willingness of data holders to share genetic data in the simplest technical context – query for the presence of a specified nucleotide at a given position (an allele) within a chromosome" from any human individual in a group (e.g., with a certain disease).
  - >200 projects are participating the Beacon project to share their human genomic data
  - Shringarpure and Bustameante recently proposed an inference attack, showing that given a an individual's whole genome sequence, an adversary may infer the presence of the individual in a beacon through repeated queries for variants in the individual's genome.
- Challenge: Given a sample Beacon database, we challenge each participating team to develop a solution to mitigate the Shringarpure-Bustamante attack, while responding a mixiumum number of queries.
  - Each team should prepare a program that responds to variation queries to any Beacon.
  - The evaluation team will evaluate the submitted programs using a Beacon that was NOT shared with the participating teams.

TRACK 1: EVALUATION CRITERIA

- General criterion: the maximum number of correct queries that an algorithm can respond before any individual in the beacon can be re-identified by the Bustamante attack.
- Procedure: we perform a (modified) Shringarpure-Bustamante attack on a beacon consisting of 500 genomes extracted from the 1000 Genomes project, through the responses from each submitted program to the queries of randomly sampled variations in the Beascon.
  - We recorded the number of correct responses (and neglected incorrect responses) until the attack power reaches 0.6.
  - The error rate is computed as: # of correct responses / total # of queries
  - The (modified) Shringarpure-Bustamante attack utilizes allele frequencies derived from the 1000 genomes project instead of those following a presumed distribution of allele frequencies
  - Only the variations in the Beacon were queried because variaions not in the database contibute little identification power for Bustamante attack

BASELINE PERFORMANCE OF TRACK 1

- Mask k% rare SNPs the database
  - Error rate: 0.2
  - Attack power reaches 0.6 when 40,000 queries perform
  - Correctly answered queries: 32,000
  - Error rate: 0.18
  - Attack power reaches 0.6 when 10,000 queries perform
  - Correctly answered queries: 8,200



TRACK 2: PRIVACY-PRESERVING SEARCH OF SIMILAR CANCER PATIENTS ACROSS ORGANIZATIONS

- Background: We consider a secure collaboration project involving two biomedical institutions: one institution hosts a sequence database of the same gene from multiple patients, and the other institution has the sequence of the gene from a single patient and wants to search it against the database to identify the patients with the top-k most similar sequences (k is typically small, <5). However, each of these two institutions cannot release their sequence data to the other institution.
  - The gene is highly divergent among different human individuals (with 85%-95% sequence identity, e.g., the immune relevant genes).
  - The sequence similarity is measured by the edit distance between a query sequence and sequences in the database. We assume the typical Secure Multiparty Computation (SMC) scenario: *no information should be leaked during the computation, except the final result*.
- Challenge: Given a gene sequence database (on Party A) and a query sequence (on Party), we challenge each participating team to develop a two-party computation algorithm to identify the top-k most similar sequences in the database.
  - The algorithm should consist of two programs, each executed on a computer of one party.
  - The algorithm should meet the securiry guarantee of SMC.
  - Approximation algorithms are allowed.

TRACK 2: EVALUATION CRITERIA

- General criterion: 1) security guarantee:the algorithms shoud not leak information other than the final results; 2) accuracy: the algorithm should report the correct top-k genes in most cases; 3) speed: the algorithm should run fast in a real-world environment, consiering both computationa and communication costs.
- Procedure: We evaluate the description of the algorithm submtted by each team; the algorithms leaking information other than the final results are disqualified. We then tested each qualified algorithms on a query gene (on one party) against a database consisting of 500 genes, in attempt to identify k=1, 3 and 5, respectively, most similar genes in the database. The ZNF717 (of ~3470 bps encoding a BRAB zinc-finger protein) gene sequences were used in the testing.
  - The submitted algorithms were executed on two virtual machines set at Indiana University and UCSD, respectively.
  - We repeated the experiment multiple times on several different databases, and recorded their running time and accuracy.
  - The algorithms are ranked according to 1) first their accuracy and 2) their running time.

TRACK 3: TESTING FOR GENETIC DISEASES ON ENCRYPTED GENOMES

- Background: We consider a secure outsourcing scenario where an biomedical institution hopes to outsource the storage and computation (in this case the search of disease markers) of human genomic data on a public cloud. The genomic data will be stored in encrypted form on the cloud, and thus the search needs to be conducted by using a homomorphic encryption protocol.
- Challenge: Given a single or multiple human genomes (in VCF format) and a genetic marker consisting a small number (<5) of variations , we challenge each participating team to develop a homomorphic encryption algorithm to encrypt the human genomes, and to test if any human genome carries the marker (i.e., containing all the variations).
  - The algorithm should consist of two programs, one for the encrytion (executed on a private computer at the biomedical institution) and one for the search (executed on the public cloud).
  - The algorithm should meet the securiry guarantee of homomorphic encryption, no other information is leaked other than the final result.

TRACK 3: EVALUATION CRITERIA

- Hide data, query and access patterns from the cloud;
- Employ homomorphic encryption;
- 80bits security;
- 1 round query/reply;
- Maximum of 5 million variants per VCF file;
- Retrieve/reveal less than 20 variants during each search;
- Maximum of 100 client-side comparison
- Maximum of 200 VCF files (number of patients).

- Client-Server model (resembling a cloud DB);
- 10Mbps network link;

- Evaluation priority
  - Speed
  - Storage
  - Communication

EVALUATION TEAMS

- Track 1: Diyue Bu (Indiana University)
- Track 2: Lei Wang, Wenhao Wang, Diyue Bu (Indiana University)
- Track 3: Chao Jiang, Feng Chen, Shuang Wang, Le Trieu Phong, Xiaoqian Jiang (UCSD)

TRACK 1: PARTICIPATING TEAMS

Team(affiliation)	Member(s)
Vanderbilt University	Zhiyu Wan Brad Malin
University of Manitoba Iran University of Science and Technology	Md Momin Al Aziz Reza Ghasemi Md Waliullah Noman Mohammed

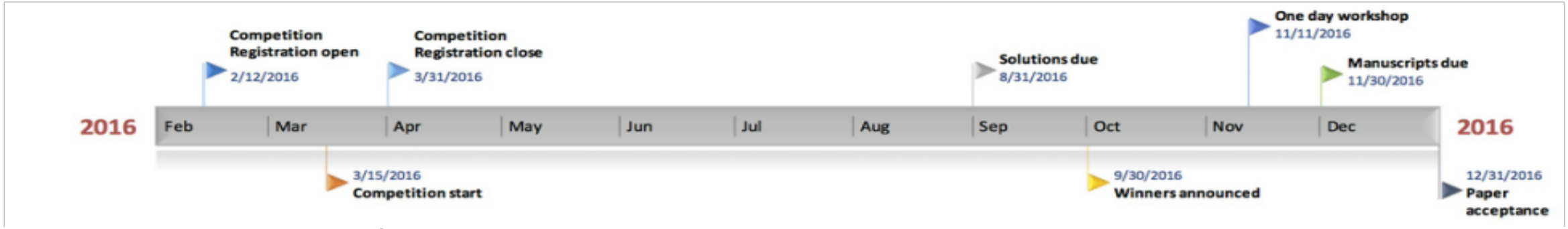
TRACK 2: PARTICIPATING TEAMS

Team(affiliation)	Member(s)	Team(affiliation)	Member(s)
IBM T.J. Watson Research Center and Bar-Ilan University, Israel.	Gilad Assharov, Shai Halevi, Yehuda Lindell, Tal Rabin	Texas A and M University	Parisa Kaghazgaran Hassan Takabi
University of Manitoba and Zayed University	Md Momin Al Aziz, Dima Alhadidi, Noman Mohammed	University of Texas at Dallas	Aref Asvadishirehjini
University of Maryland	Xiao Wang, Jonathan Katz	Cybernetica AS	Dan Bogdanov Peeter Laud Ville Sokk Sander Siim
Indiana University, Bloomington	Ruiyu Zhu, Yan Huang	Communication and Distributed Systems, RWTH Aachen University	

TRACK 3: PARTICIPATING TEAMS

Team (affiliation)	Member(s)	Team (affiliation)	Member(s)
Microsoft research	Kristin Lauter, Kim Laine, Hao Chen, Gizem Cetin, Peter Rindal, Yuhou (Susan) Xia	IBM	Hamish Hunt, Flavio Bergamaschi, Shai Halevi
Communication and Distributed Systems, RWTH Aachen University, Germany	David Hellmanns, Martin, Henze, Jens Hiller, Ike Kunze, Sven Linden, Roman Matzutt, Jan Metzke, Marco Moscher, Jan Pennekamp, Felix Schwinger, Klaus Wehrle, Jan Henrik Ziegeldorf	EPFL team	João Sá Sousa, Cédric Lefebvre, Zhicong Huang, Jean Louis Raisaro, Florian Tramer, Carlos Aguilar, Jean-Pierre Hubaux, Marc-Olivier Killijian
Waseda University	Yu Ishimaki Hayato Yamana	University of Texas at Dallas	Ehsan Hesamifard
Seoul National University	Jung Hee Cheon, Miran Kim, Yongsoo Song		

WORKSHOP PREPARATION AND REGISTRATION



Registered Teams

- 13 countries
- 50+ teams

EVALUATION RESULTS

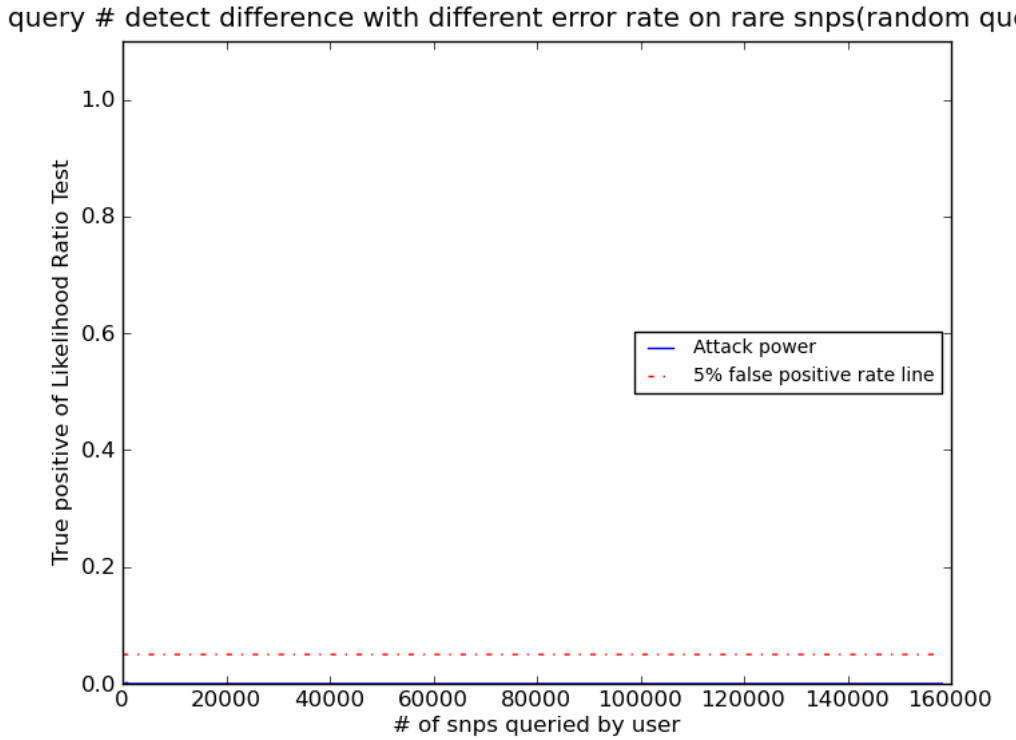
BEST-PERFORMING TEAMS & RESULTS

-- Result displayed is the best performance among team's submission of mitigation methods

- Team: Zhiyu Wan (Vanderbilt University)

  Brad Malin (Vanderbilt University)

- Result: No power presents even when 160,000 queries performed
- Error rate: 0.115
- Correctly answered queries: 141,600

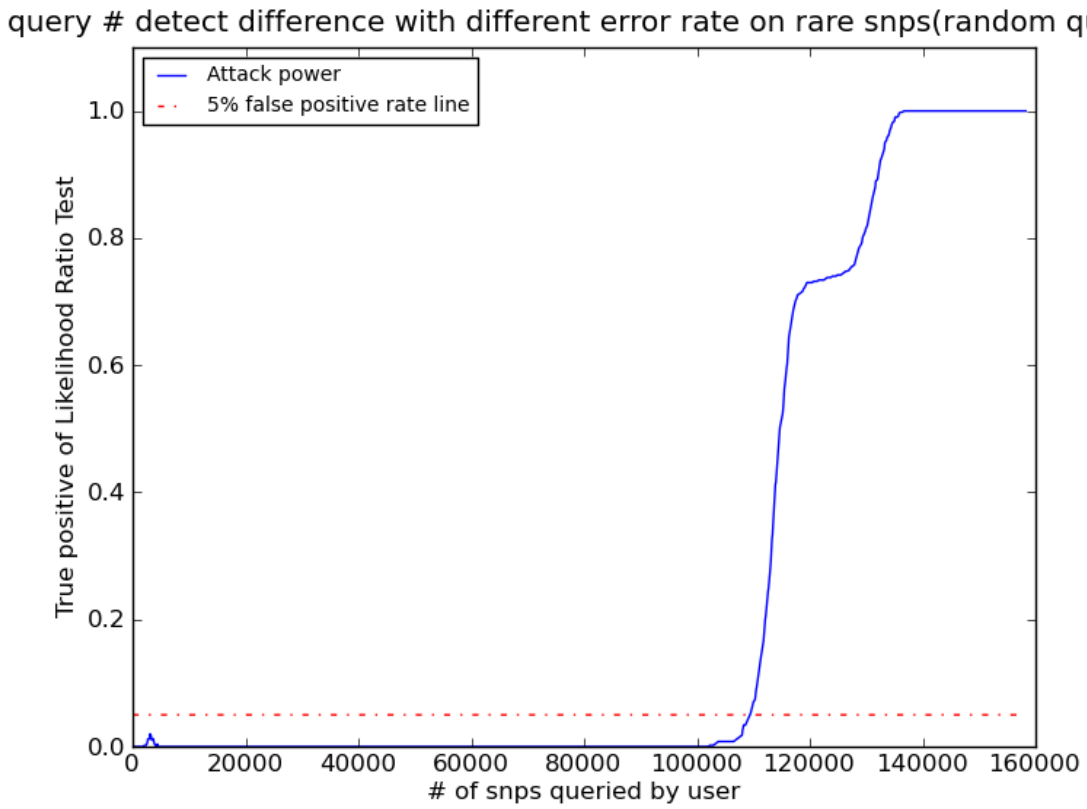


BEST-PERFORMING TEAMS & RESULTS

-- Result displayed is the best performance among team's submission of mitigation methods

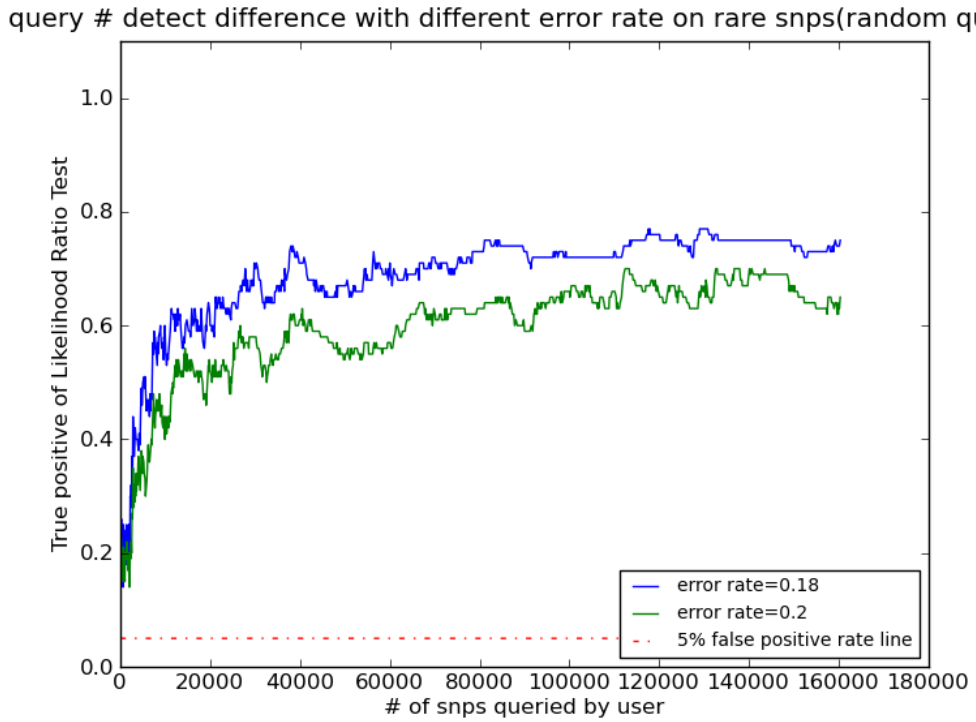
- **Team:** Md Momin Al Aziz (University of Manitoba)
  Reza Ghasemi (Iran University of Science and Technology)
  Md Waliullah (University of Manitoba)
  Noman Mohammed (University of Manitoba)

- Result attack power reaches 0.6 when around 110,000 queries performed:
- Error rate: 0.509
- Correctly answered queries: 54,010



BASELINE PERFORMANCE OF TRACK 1

- Mask k% rare SNPs the database
  - Error rate: 0.2
  - Attack power reaches 0.6 when 40,000 queries perform
  - Correctly answered queries: 32,000
  - Error rate: 0.18
  - Attack power reaches 0.6 when 10,000 queries perform
  - Correctly answered queries: 8,200



TRACK 2: BEST-PERFORMING TEAMS

-- Evaluated by database with 500 patients records, run-time shown as average ± std through 5 runs

Team	Members	Top 1 Run-time(s)	Top 1 Accuracy	Top 3 Run-time(s)	Top 3 Accuracy	Top 5 Run-time(s)	Top 5 Accuracy
IBM T.J. Watson Research Center and Bar-Ilan University, Israel.	**Gilad Assharov, Shai Halevi, Yehuda Lindell, Tal Rabin**	**11.37 ±0.31**	**correct**	**11.41 ±0.17**	**2 or 3 correct**	**11.62 ±0.38**	**4 or 5 correct**
University of Manitoba and Zayed University	Md Momin Al Aziz, Dima Alhadidi, Noman Mohammed	22.65 ±0.11	Not correct	22.99 ±0.15	2 correct	22.88 ±0.37	3 correct
University of Maryland	Xiao Wang, Jonathan Katz	12.93 ±1.26	correct	21 ±0.9	1 or 2 correct	30.4 ±2.93	2 or 3 correct
Indiana University, Bloomington	Ruiyu Zhu, Yan Huang	209.03 ±7.58	correct	273.14 ±7.02	All correct	337.79 ±6.18	4 or 5 correct

TRACK 2: BEST-PERFORMING TEAMS

-- Evaluated by database with 500 patients records, run-time shown as average ± std through 5 runs

Team	Top 1 Run-time(s)	Top 1 Accuracy	Top 3 Run-time	Top 3 Accuracy	Top 5 Run-time(s)	Top 5 Accuracy
Texas A and M University **(Dis-Qualified)**	235.19 ±5.75	correct	335.50 ±14.65	All correct	525.29 ±5.08	All correct
University of Texas at Dallas	64.74	Not correct	68.72	Not correct	98	Not correct
Cybernetica AS	80.97 ±33.45	correct	67.47 ±5.39	1 correct	64.64 ±5.85	1 correct
RWTH Aachen University	95m	correct	>105m	All correct	>105m	All correct

TRACK 3: QUERIES AND DATASETS

- 1 query (4 variants) vs. 1 VCF file [10K records]
- 1 query (4 variants) vs. 1 VCF file [100K records]
- 1 query (1 variant) vs. 50 VCF files [100K records]

EXPERIMENTAL RESULTS (1 QUERY (4 VARIANTS) / 1 VCF [10K])

Teams	The setup time [including key generation, database encryption, and upload] (s)	Size of the encrypted DB (MB)	Time to compare the query and the encrypted DB (s)	Memory usage of the server (MB)	Time to decrypt the results (s)	Size of the encrypted results (MB)	Total turnaround time [compare + transfer + decrypt] (s)	Rank
Microsoft	0.84	3.8	0.56	80	0.025	0.644	1.10	1
SNU	47.41	4.0	8.49	164	0.002	2	10.09	2
COMSYS	32.35	255.0	15.16	90	0.670	0.434	16.18	3
EPFL	137.03	146.8	6.846	386	9.366	3.998	19.41	4
NTU	619.94	1242.0	55485.6	1790	0.600	1.2	55487.20	7
IBM	538.29	1660.0	1177.59	3807	230.250	23	1426.24	6
WU	40.92	549.9	933.771	1448	0.061	1.558	935.08	5

EXPERIMENTAL RESULTS (1 QUERY (4 VARIANTS) / 1 VCF [100K])

Teams	The setup time [including key generation, database encryption, and upload] (s)	Size of the encrypted DB (MB)	Time to compare the query and the encrypted DB (s)	Memory usage of the server (MB)	Time to decrypt the results (s)	Size of the encrypted results (MB)	Total turnaround time [compare + transfer + decrypt] (s)	Rank
Microsoft	1.86	24	3.09	224	0.024	0.644	3.6292	1
SNU	51.02	10	21.1003	340	0.00495	5	25.11	4
COMSYS	34.9	255	15.28	90	0.68	0.444	16.32	2
EPFL	137.6	147	6.79	3846	9.28	3.99	19.26	3
NTU	n/a	n/a	n/a	n/a	n/a	n/a	n/a	7
IBM	478.1	1660	959.1	3713	200.7	23	1178.2	5
WU	109.721	5447.82	8937.51	2779	0.05776	1.56	8938.81	6

EXPERIMENTAL RESULTS (1 QUERY (1 VARIANT) / 50 VCF [10K])

Teams	The setup time [including key generation, database encryption, and upload] (s)	Size of the encrypted DB (MB)	Time to compare the query and the encrypted DB (s)	Memory usage of the server (MB)	Time to decrypt the results (s)	Size of the encrypted results (MB)	Total turnaround time [compare + transfer + decrypt] (s)	Rank
Microsoft	36.69	188	32.77	83.6	1.21	32	59.58	1
SNU	2384	244	129.28	85.70	0.03218	122	226.9	2
COMSYS	1207.07	13000	278.81	96.264	0.79	22	297.2	4
EPFL	6903.1	1468	122	3855	127	49.97	288.9	3
NTU	n/a	n/a	n/a	n/a	n/a	n/a	n/a	6
IBM	n/a	n/a	n/a	n/a	n/a	n/a	n/a	6
WU	2102.28	27491.11	12447.98	72003.1	23.17	77.92	12533.48	5

TASK3: BEST PERFORMING TEAMS

Teams	1 query (4 variants) / 1 VCF [10k]		1 query (4 variants) / 1 VCF [100k]		1 query (1 variant) / 50 VCF [100k]		Overall score	
	Total turnaround time (s)	Rank	Total turnaround time (s)	Rank	Total turnaround time (s)	Rank		
Microsoft	1.10	1	3.6292	1	59.58	1	**1**	Winner
SNU	10.09	2	25.11	4	226.9	2	**2.67**	
COMSYS	16.18	3	16.18	2	297.2	4	**3**	Runner-up
EPFL	19.41	4	19.26	3	288.9	3	**3.33**	

CALL FOR FULL PAPER SUBMISSION

- Special issue in 

BMC Medical Genomics

- Peer-review
- Submission deadline: Dec-31-2016
- Notice of decision: Jan-31-2017

- http://humangenomeprivacy.org/2016/paperSubmission

ACKNOWLEDGEMENT

- Human Longevity Inc. and GeneCloud for providing cash awards.

- NIH grants (U54HL108460, R13HG00907201A1) to support the competition

Thanks for the Participation

LUNCH BREAK

PRESENTATIONS FROM PARTICIPATION TEAMS: TRACK 1

- Zhiyu Wan, Brad Malin, (Vanderbilt University)

- Md Momin Al Aziz (University of Manitoba), Reza Ghasemi (Iran University of Science and Technology), Md Waliullah, Noman Mohammed, (University of Manitoba)

PRESENTATIONS FROM PARTICIPATION TEAMS: TRACK 2

- Gilad Asharov (Cornell), Shai Halevi (IBM), Yehuda Lindell (Bar-Ilan University), Tal Rabin (IBM)

- Md Momin Al Aziz, Dima Alhadidi*, Noman Mohammed, (University of Manitoba, *Zayed University)

- Xiao Wang, Jonathan Katz, (University of Maryland)

- Ruiyu Zhu, Yan Huang, (Indiana University, Bloomington)

PRESENTATIONS FROM PARTICIPATION TEAMS: TRACK 3

- Kristin Lauter, Kim Laine, Hao Chen, (Microsoft research), Gizem Cetin, (Worcester Polytechnic Institute), Peter Rindal, (Oregon State University), Yuhou (Susan) Xia, (Princeton University)
- Jung Hee Cheon, Miran Kim, Yongsoo Song, (Seoul National University)
- David Hellmanns, Martin, Henze, Jens Hiller, Ike Kunze, Sven Linden, Roman Matzutt, Jan Metzke, Marco Moscher, Jan Pennekamp, Felix Schwinger, Klaus Wehrle, Jan Henrik Ziegeldorf, (RWTH Aachen University, Germany)
- João Sá Sousa, (EPFL) Cédric Lefebvre, (Université Toulouse), Zhicong Huang, Jean Louis Raisaro, Florian Tramer, (EPFL) Carlos Aguilar, (Université Toulouse), Jean-Pierre Hubaux, (EPFL), Marc-Olivier Killijian, (Université Toulouse)