Privacy in the Genomic Era

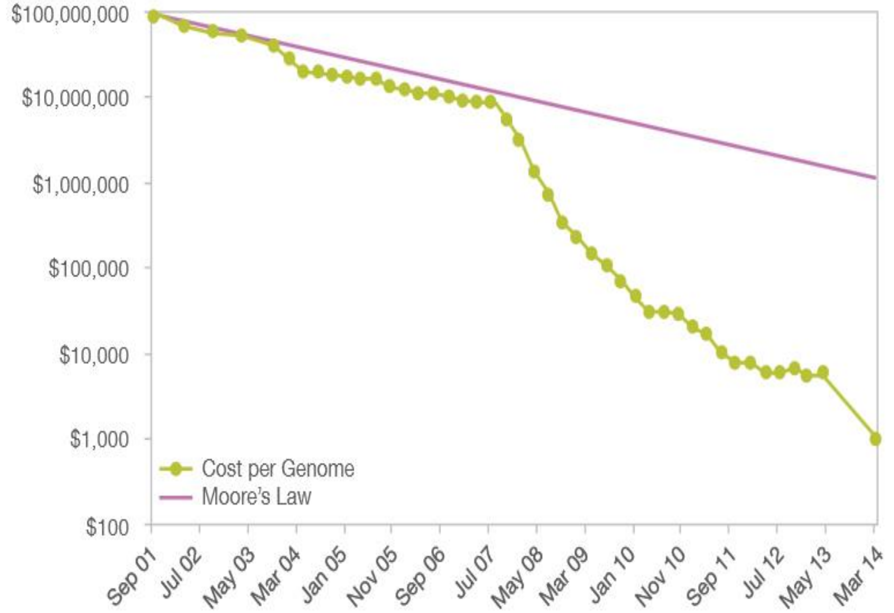
XiaoFeng Wang, IUB

http://www.informatics.indiana.edu/xw7

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

  ……

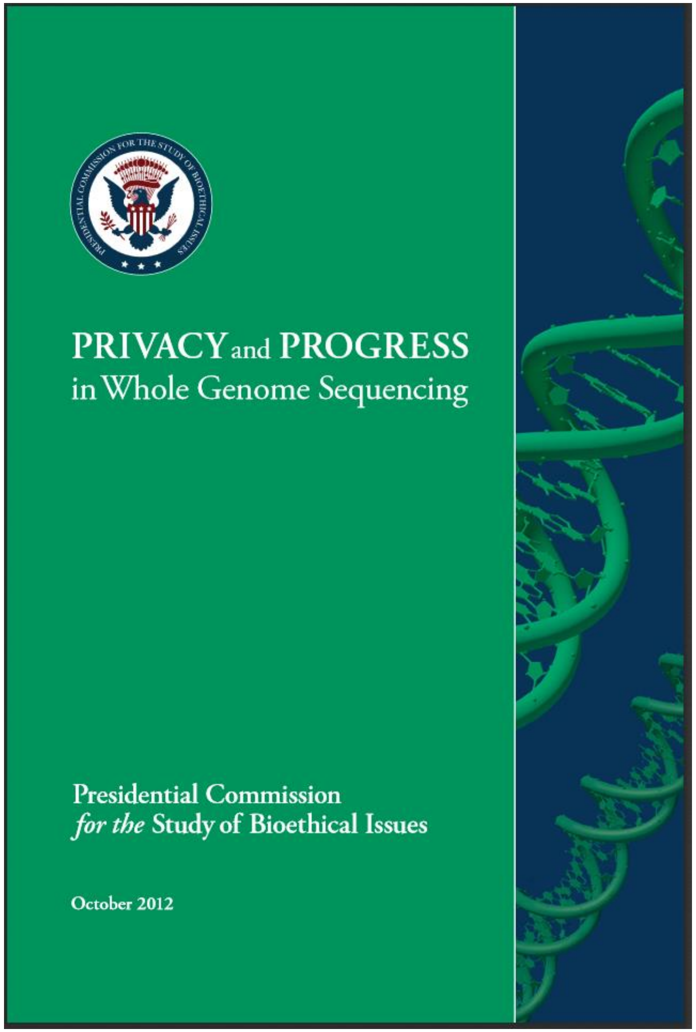


Genome Privacy

- 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

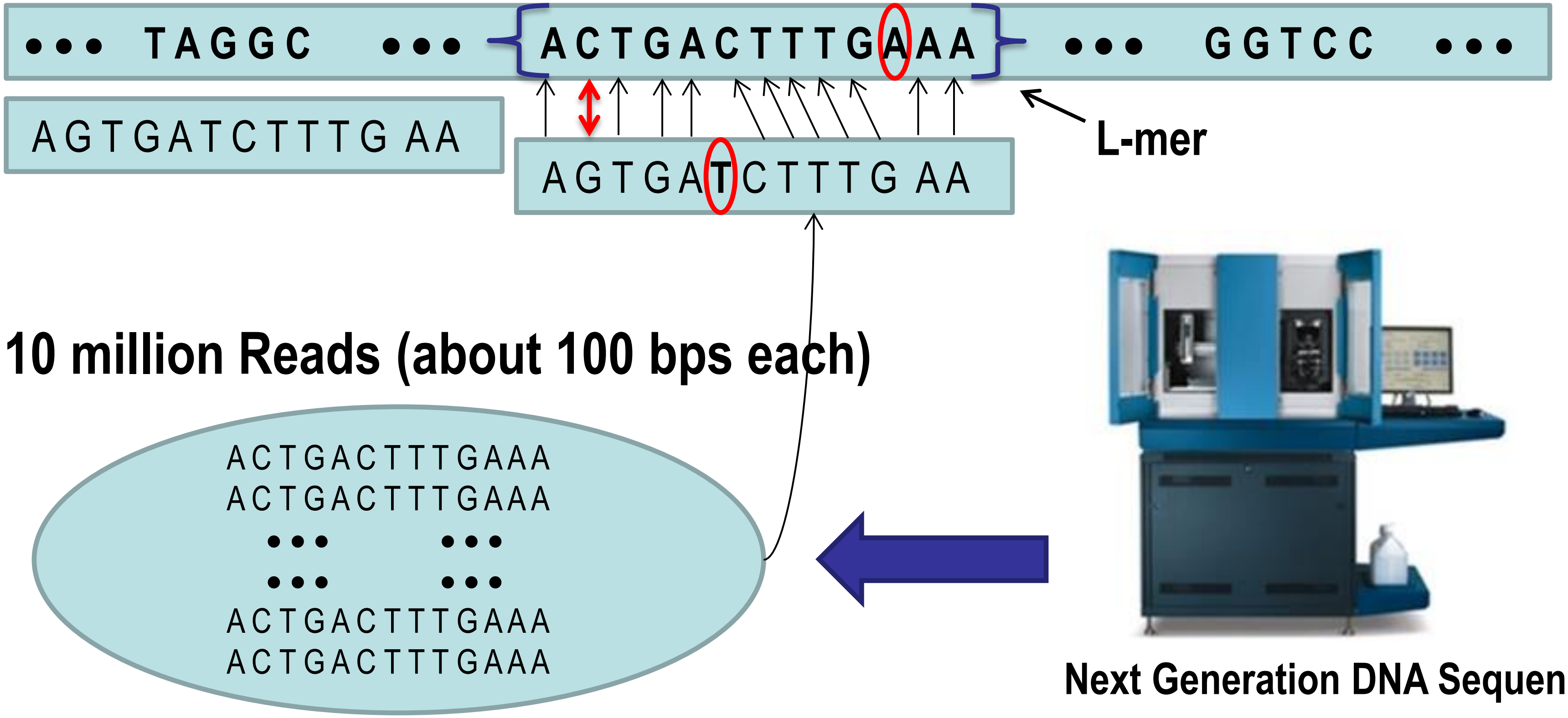
Available at http://arxiv.org/pdf/1405.1891v1.pdf

Technical Challenges

- Dissemination: anonymization is difficult !
  - Extremely high dimensions
  - Hard to balance between privacy and utility
- Computing: big data analysis
  - Beyond the capability of existing secure computing technologies

Secure Elastic Read Mapping and Filtering

Reference Genome (about 6 billion bps for two strands)



••• TAGGC ••• ACTGACTTTGAAA ••• GGTCC •••

AGTGATCTTTGAA

AGTGATCTTTGAA

L-mer

10 million Reads (about 100 bps each)

ACTGACTTTGAAA
ACTGACTTTGAAA
••• •••
••• •••
ACTGACTTTGAAA
ACTGACTTTGAAA

Next Generation DNA Sequencer

Big Data Analysis

- Technical Challenges
  - ➢ Millions of reads and a reference of billions of nucleotides
  - ➢ Edit-distance based alignment
- Cloud solutions
  - ➢ Cost of sequencing < cost of mapping within organizations
  - ➢ Cloud computing is the only solution
- Privacy
  - ➢ NIH disallows reads with human DNA to be given to the public Cloud

Privacy-preserving Genomic Data Sharing

- Old problems:
  - Statistical inference control, access control, query auditing…
- However, genome data are special:
  - Special structures, e.g. linkage disequilibrium
  - Existence of reference genomic data that are publicly available (e.g. large population studies as HapMap, WTCCC, 1000 Genome)
- An example: Homer’s attack and NIH’s responses

Our Research

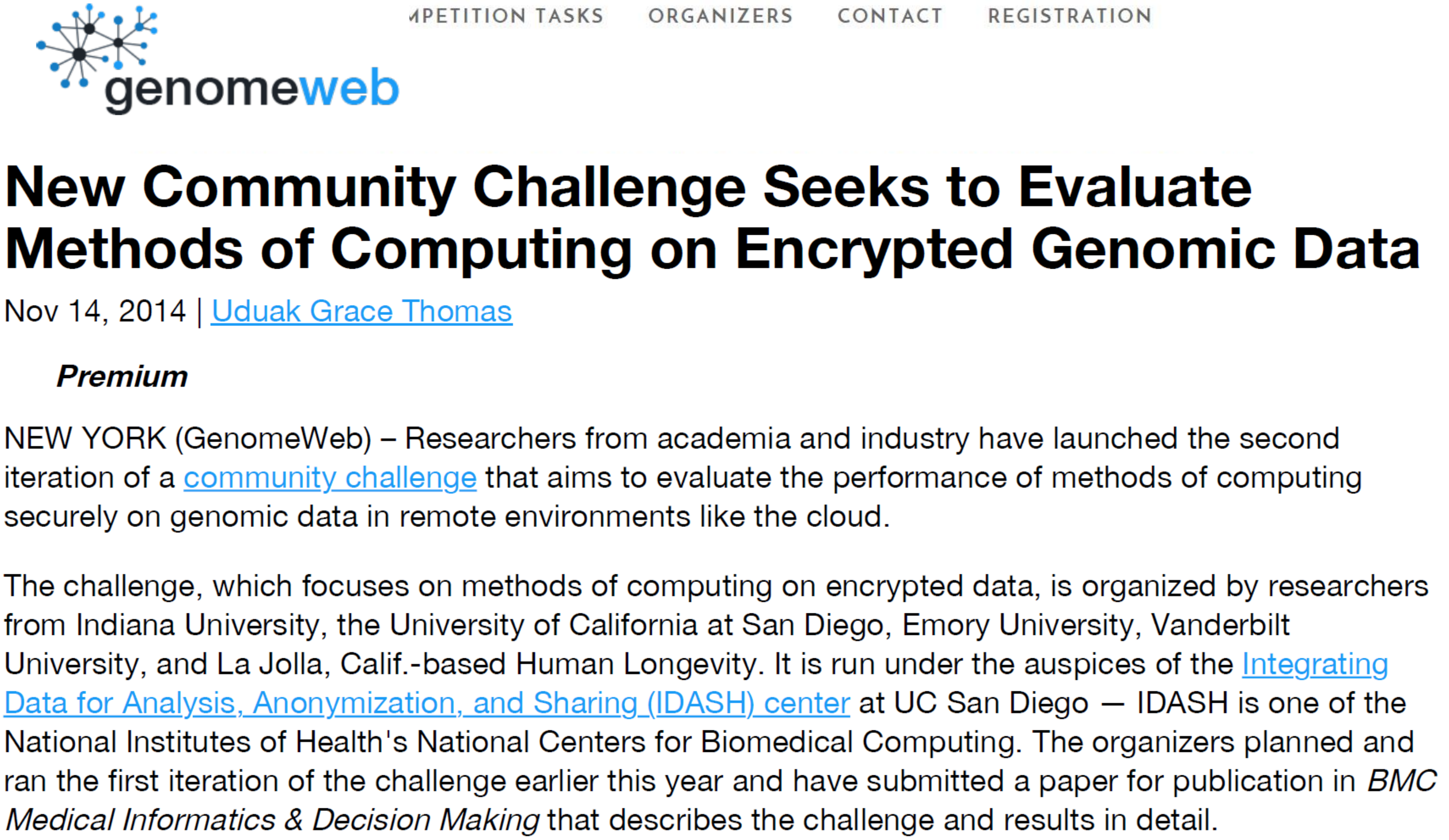
- Our prior discovery: ID from GWAS publications
  - Test statistics → Allele Frequencies
  - LD statistics → Statistical Identification
  - Pair-wise allele frequencies → SNP Sequences
- **Research on the risk advisory system for genome data sharing**
  - **Red (risky), Yellow (potentially risky), Green (safe)**

- **Research on DNA data protection**
  - **Balance between risk mitigation and data utility**

For More Information

1. Choosing Blindly but Wisely: Differentially Private Solicitation of DNA Datasets for Disease Marker Discovery 2014 JAMIA
2. Large-Scale Privacy-Preserving Mappings of Human Genomic Sequences on Hybrid Clouds 2012 NDSS
3. To Release or Not to Release: Evaluating Information Leaks in Aggregate Human-Genome Data 2011 ESORICS
4. Learning Your Identity and Disease from Research Papers: Information Leaks in Genome Wide Association Study 2008 CCS

Community Challenges on Genome Privacy !



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.

Challenge 2014

- **Theme**: Genome Data Anonymization and Sharing
  - Protecting SNP sequences: 200 individuals, 311 to 610 SNPs
  - Protecting GWAS results: 201 cases/174 controls, 5000 to 106,129 SNPs
- **Participants**:
  - U Oklahoma, UT Dallas, McGill, UT Austin and CMU
- **Outcomes**: evaluated by a biomedical and security panel
  - Great promising for sharing GWAS results: Austin won the competition
  - Difficulty in sharing raw data: existing techniques cannot preserve data utility

Challenge 2015 !

- **Objective:**

Find out how close secure computing technologies are in supporting real-world genomic data analysis

- **Challenges:**
  - Secure outsourcing: HME-based analysis on encrypted genome sequences (GWAS analysis, sequence comparison)
  - Secure collaboration: SMC-based data analysis across the Internet

- **Deadline:**
  - Registration is now open
  - Deadline for submitting the result (code): March 1st.
  - Workshop: March 16 at UCSD

HOW to PARTICIPATE

Goto:

http://www.humangenomeprivacy.org

Acknowledge

- NIH R01 (1R01HG007078-01): “Privacy Preserving Technologies for Human Genome Data Analysis and Dissemination”

- NSF-CNS-1408874: “Broker Leads for Privacy-Preserving Discovery in Health Information Exchange”