Proposed Social and Technological Solutions to Issues of Data Privacy in Personal Genomics

(Licensure, Secondary Datasets, Enclaves)

Mark Gerstein
Yale

Slides freely downloadable from Lectures.GersteinLab.org & “tweetable” (via @markgerstein). See last slide for references & more info.
Setting the Stage: 
the Advent of Personal Genomics

• Human Genome sequence in 2000 for ~$3 billion
• A Human Genome can be sequenced today for ~$1000
• Thousands of SNPs can be interrogated for ~$99
The Explosion of Data in Genomics: the Numbers

From ‘00 to ~’ 20, cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

[Nature 507, 294; Sboner et al. (‘11) GenomeBiology]
DTC Genomics

• Industry spurred by falling prices of sequencing and computation
• Major players were Navigenics, DeCode and 23andMe.
• 23andMe
  – has 600,000 Customers
  – $99 per analyiss
  – Promotes sharing of Data
  – Currently in trouble with the FDA and limited to only recreational (e.g., ancestry related) analysis
  – Millions in VC funding
Proposed Social & Technological Solutions to Issues of Data Privacy in Personal Genomics (Licensure, Secondary Datasets, Enclaves)

• Setting the Stage: the Advent of Personal Genomics
• The of Conundrum Genomic Privacy
  – Fundamental, inherited info v need for data-sharing to enable research

• Current Social & Tech Approaches
  – GINA, Consents & "Secure" use of dbGAP
  – Ways the solutions have been "hacked"

• Strawman Hybrid Soc-Tech Approach
  – Soc: Licensure for genomic researchers
  – Enclaves in the cloud
  – Intelligent data formats mostly splitting private & public info (ex of MRF)
  – Small Amounts of leakage
Privacy is a personal and fundamental right guaranteed by the US Constitution
Privacy Act 1974

Including:
• Inherent in the limits on the First Amendment is a constitutional right to privacy.
• Fourth Amendment against search and seizure
  
  US v Amerson 483 F. 3d 73 (2d Cir. 2007);
• Due Process Clauses of the Fifth and Fifteenth Amendments.
The Conundrum of Genomic Privacy: Is it a Problem?

Yes
Genetic Exceptionalism
Not yet sure of the relevance of the data (but the internet doesn’t forget)
Testing discloses both yours’ and your family’s fundamental data

No
Shifting societal foci
No one really cares about your genes
You might not care
Cost Benefit Analysis: how helpful is identifiable data in genomic research?

[Klitzman & Sweeney ('11), J Genet Couns 20:981; Greenbaum & Gerstein ('09), New Sci. (Sep 23)]
Tricky Privacy Considerations in Personal Genomics

• Personal Genomic info. essentially meaningless currently but will it be in 20 yrs? 50 yrs?
  – Genomic sequence very revealing about one’s children
  – Once put on the web it can’t be taken back

• Not clear whether they can be treated with “open data” ethos of circa 2000

• Ownership of the data & what consent means (Hela)
  – Could your genetic data give rise to a product line?

• Large discussion of Identification Risk but what about Characterization Risk
  – Finding you were in study X vs identifying that you have trait Y from studying your identified genome

[D Greenbaum & M Gerstein ('08). Am J. Bioethics; D Greenbaum & M Gerstein, Hartford Courant, 10 Jul. '08 ; SF Chronicle, 2 Nov. '08; Greenbaum et al. PLOS CB ('11); Greenbaum & Gerstein ('13), The Scientist; Photo from NY Times]
The Other Side of the Coin: Why we should share

• Sharing helps speed research
  – Large-scale mining of this information is important for medical research
  – Privacy is cumbersome, particularly for big data
  – Sharing is important for reproducible research
  – Sharing is useful for education
• The individual (harmed?) v the collective (benefits)
  – But do sick patients care about their privacy?
• What is acceptable risk? What is acceptable data leakage?
• Maybe a we need a few "test pilots" (ala PGP)?
  – Sports stars & celebrities?

Genomics has similar "Big Data" Issues to the Rest of Society

- Sharing & "peer-production" is central to success of many new ventures, with the same risks as in genomics
- We confront privacy risks every day we access the internet
- (...or is the genome more exceptional & fundamental?)

[Seringhaus & Gerstein ('09), Hart. Courant (Jun 5); Greenbaum & Gerstein ('11). Times (6 Oct)]
Proposed Social & Technological Solutions to Issues of Data Privacy in Personal Genomics (Licensure, Secondary Datasets, Enclaves)

• Setting the Stage: the Advent of Personal Genomics
• The of Conundrum Genomic Privacy
  – Fundamental, inherited info v need for data-sharing to enable research

• Current Social & Tech Approaches
  – GINA, Consents & "Secure" use of dbGAP
  – Ways the solutions have been "hacked"

• Strawman Hybrid Soc-Tech Approach
  – Soc: Licensure for genomic researchers
  – Enclaves in the cloud
  – Intelligent data formats mostly splitting private & public info (ex of MRF)
  – Small Amounts of leakage
Health Insurance Portability and Accountability Act of 1996

• protects individuals from being charged higher premiums based on Genetics
• does not protect groups from being charged higher premiums
• Treats Genetic information like all other health information:
  – to be protected it must meet the definition of protected health information (PHI):
    • it must be individually identifiable
    • and maintained by a covered health care provider, health plan, or health care clearinghouse.
    • See 45 C.F.R 160.103 and 164.501
  – a use or disclosure of genetic information in violation of the HIPAA Privacy Rule could result in a fine of $100 to $50,000 or more for each violation.
• HOWEVER
• the regulation does not address the type of information that is protected but, rather, who holds it
• many facilities that perform direct-to-consumer genetic testing and analysis are exempt
• HIPAA Doesn’t prohibit
  – Requiring or requesting genetic tests
  – Disclosure of genetic data without permission
  – Excluding coverage for a condition
• Anonymized biological material is not considered PHI
Genetic Information Nondisclosure Act of 2008

• Title I relating to Health Insurance
• Title II relating to Employment Discrimination
• GINA Prohibits:
  – group and individual health insurers from using genetic data for determining eligibility or premiums
  – insurers from requesting that the insured undergo genetic testing
  – employers from using genetic data to may employment decisions
  – Employers from requesting genetic data about an employee or their family
Current Social & Technical Solutions

- Consents
- dbGAP distribution of data
- Local computes on secure computer
- Issues
  - Non-uniformity of consents & paperwork
  - Encryption & computer security creates burdensome requirements on data sharing & large scale analysis
    - Security increasingly becoming harder & harder
  - Different international norms

[Smith et al ('05), Genome Biol.; Greenbaum et al ('04), Nat. Biotech; Greenbaum & Gerstein ('13), The Scientist]
Proposed Social & Technological Solutions to Issues of Data Privacy in Personal Genomics (Licensure, Secondary Datasets, Enclaves)

• Setting the Stage: the Advent of Personal Genomics

• The of **Conundrum Genomic Privacy**
  – Fundamental, inherited info v need for data-sharing to enable research

• Current Social & Tech Approaches
  – GINA, Consents & "Secure" use of dbGAP
  – Ways the solutions have been "hacked"

• **Strawman Hybrid Soc-Tech Approach**
  – Soc: Licensure for genomic researchers
  – Enclaves in the cloud
  – Intelligent data formats mostly splitting private & public info (ex of MRF)
  – Small Amounts of leakage
Matching against reference genotype. The number of DNA markers such as single-nucleotide polymorphisms (SNPs) that are needed to uniquely identify a single person is small; Lin et al. estimate that only 30 to 80 SNPs could be sufficient.

Linking to nongenetic databases. A second route to identifying genotyped subjects is deduction by linking and then matching geno-type-plus-associated data (such as gender, age, or disease being studied) with data in health-care, administrative, criminal, disaster response, or other databases … If the nongenetic data are overtly identified, the task is straightforward.

- a framework for accurately and robustly resolving whether individuals are in a complex genomic DNA mixture using high-density single nucleotide polymorphism (SNP) genotyping microarrays.
- We demonstrate an approach for rapidly and sensitively determining whether a trace amount (<1%) of genomic DNA from an individual is present within a complex DNA mixture.
  - Identifying anonymized 1000G individuals through DB xref
About the James Watson Genotype Viewer

On May 31, 2007, Nobel Laureate James Watson received his personal genome sequence at a ceremony at the Baylor College of Medicine. This genome sequence describes the six billion base pairs of DNA that James Watson received from his two parents, the unique combination of which are responsible for James Watson's genetic individuality. Dr. Watson is making his genome sequence available to the public in the hope that it will encourage the development of an era of "personalized medicine" when the information contained in our genomes is used to identify and prevent diseases to which we are genetically prone before they appear, and to create personalized medical therapies that have the maximum benefit and the minimum risk. This simple browser allows you to view the places where Watson's sequence is different from the "reference" human genome sequence, as well as to view the genes and some of the common diseases associated with them.

What the Watson Sequence is

Dr. Watson's genome was sequenced at 6x coverage using 454 Life Sciences Technology. This means that each position on the genome was sequenced roughly six times. However, because of the probabilistic nature of the technology, some positions have been seen more than six times, and some less or not at all. The 454 technology produces short stretches of sequences called "reads" that are roughly 100 bases long. However, the functional units of the genome, the genes, are roughly 50,000 bases long, or 500 times the size of a 454 sequence. To interpret the Watson sequence, it was matched to the sequence, with the exception of the ApoE gene, variants of which are associated with Alzheimer's disease.

A variant or polymorphism is a small alteration in the DNA sequence. Each SNP has two possible alleles. Each SNP has two possible alleles.
Robust De-anonymization of Large Datasets
(How to Break Anonymity of the Netflix Prize Dataset)

Arvind Narayanan and Vitaly Shmatikov
The University of Texas at Austin
February 5, 2008

Abstract

We present a new class of statistical de-anonymization attacks against high-dimensional micro-data, such as individual preferences, recommendations, transaction records and so on. Our techniques are robust to perturbation in the data and tolerate some mistakes in the adversary's background knowledge.

We apply our de-anonymization methodology to the Netflix Prize dataset, which contains anonymous movie ratings of 500,000 subscribers of Netflix, the world's largest online movie rental service. We demonstrate that an adversary who knows only a little bit about an individual subscriber can easily identify this subscriber's record in the dataset. Using the Internet Movie Database as the source of background knowledge, we successfully identified the Netflix records of known users, uncovering their apparent political preferences and other potentially sensitive information.

Cross correlated small set of identifiable IMDB movie database rating records with large set of “anonymized” Netflix customer ratings
Proposed Social & Technological Solutions to Issues of Data Privacy in Personal Genomics
(Licensure, Secondary Datasets, Enclaves)

• Setting the Stage: the Advent of Personal Genomics

• The of **Conundrum Genomic Privacy**
  – Fundamental, inherited info v need for data-sharing to enable research

• Current Social & Tech Approaches
  – GINA, Consents & "Secure" use of dbGAP
  – Ways the solutions have been "hacked"

• **Strawman Hybrid Soc-Tech Approach**
  – Soc: Licensure for genomic researchers
  – Enclaves in the cloud
  – Intelligent data formats mostly splitting private & public info (ex of MRF)
  – Small Amounts of leakage
Strawman Hybrid Social & Tech Solution?

- Fundamentally, researchers have to keep genetic secrets
  - Genetic Licensure
- Technology to make it easier
  - Cloud computing & enclaves
  - Selection of stubb & "test pilot" datasets for benchmarking
  - Careful separation of private & public data
    - Lightweight, freely accessible secondary datasets
- Technological barriers shouldn't create a social incentive for “hacking”

Example Format for Functional Genomics
(Gene Expression Levels)

• Human genome reseq. all about variants vs. reference
• Situation diff. for func. genomics
  – Often variant info. Is determined incidentally

• On one hand: **Reads have variant information** in most functional regions (deep RNA-seq expt. essentially exome seq.)
• On other hand: **high-level summaries and signal tracks mostly what is used (80%) and do not involve variant info.** Helpful to make this freely available and easy to use

[Greenbaum et al. *PLOS CB* ('11)]
Light-weight formats

- Some lightweight format clearly separate public & private info., aiding exchange
- Files become much smaller
- Distinction between formats to compute on and those to archive with – become sharper with big data

[Bioinformatics 27: 281]


**MRF Examples**

10X Compression Ex.

Raw **ELAND** export file has uncompressed file size: ~4 GB; total number of reads: ~20 million; number of mapped reads: ~12 million.

**MRF file** is significantly smaller (~400 MB uncompressed, ~130 MB compressed with gzip).

**BAM file** has a size of ~1.2 GB.

Reference based compression (ie CRAM) is similar but it stores actual variant beyond just position of alignment block.

[Habegger et al., Bioinformatics (‘11)]
Sensitive Information Linkage

Gene Expression Levels → Age, Height, Weight, BMI, ...

Gender

Race and Population

Variants

Expression Genotype Association Databases

Phenotype Genotype Association Databases

Phenotype Associated Variants

Personally Identifying Variants

X
Small Data Leakage from just Gene Expression Data:
4 eQTL-SNP genotypes

Example: Vulnerable sample variants, expressions
- Variant 0 (1, 6)
- Variant 1 (0, 2)
- Variant 2 (1, 3)
- Variant 3 (0, 2)

Expression levels are outliers and are predictive of the genotype!
Proposed Social & Technological Solutions to Issues of Data Privacy in Personal Genomics  
(Licensure, Secondary Datasets, Enclaves)

• Setting the Stage: the Advent of Personal Genomics

• The of **Conundrum Genomic Privacy**
  - Fundamental, inherited info v need for data-sharing to enable research

• Current Social & Tech Approaches
  - GINA, Consents & "Secure" use of dbGAP
  - Ways the solutions have been "hacked"

• **Strawman Hybrid Soc-Tech Approach**
  - Soc: Licensure for genomic researchers
  - Enclaves in the cloud
  - Intelligent **data formats mostly splitting private & public info** (ex of MRF)
  - Small Amounts of leakage
Acknowledgements

papers.gersteinlab.org/subject/privacy

D Greenbaum

A Harmanci

A Sboner, XJ Mu

L Habegger, A Sboner, TA Gianouli, J Rozowsky, A Agarwal, M Snyder
Default Theme

- Default Outline Level 1
  - Level 2
More Information on this Talk

SUBJECT: Networks

DESCRIPTION:

NOTES:
This PPT should work on mac & PC. Paper references in the talk were mostly from Papers.GersteinLab.org.

PERMISSIONS: This Presentation is copyright Mark Gerstein, Yale University, 2010. Please read permissions statement at http://www.gersteinlab.org/misc/permissions.html. Feel free to use images in the talk with PROPER acknowledgement (via citation to relevant papers or link to gersteinlab.org).

PHOTOS & IMAGES. For thoughts on the source and permissions of many of the photos and clipped images in this presentation see http://streams.gerstein.info. In particular, many of the images have particular EXIF tags, such as kwopotppt, that can be easily queried from flickr, viz: http://www.flickr.com/photos/mbgmbg/tags/kwopotppt.